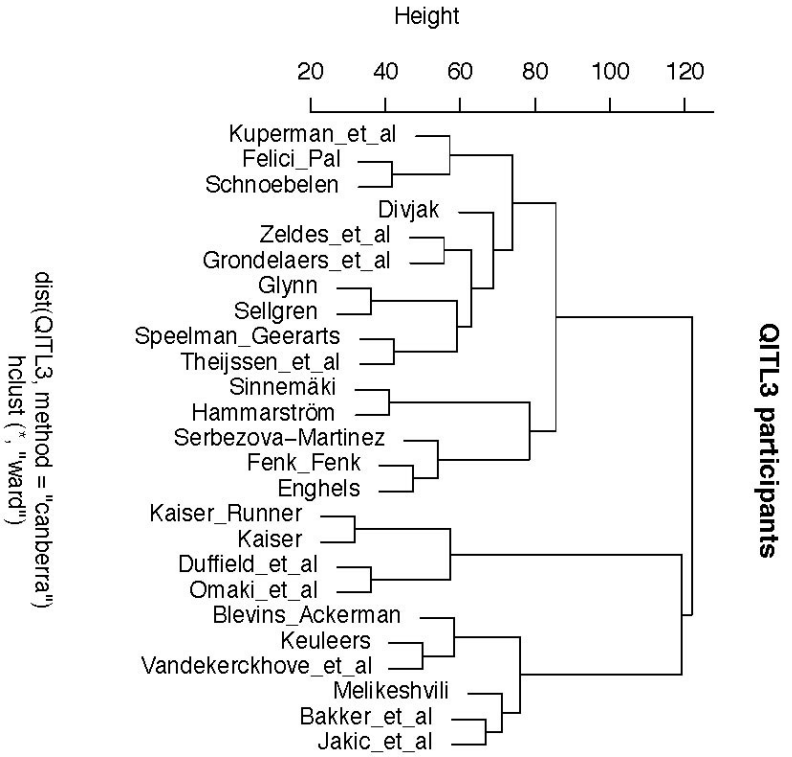
Third Workshop on
Quantitative Investigations in Theoretical Linguistics
(QITL3)



Arppe, Antti; Kaius Sinnemäki; and Urpo Nikanne
(Editors)

Third Workshop
on

Quantitative Investigations In Theoretical Linguistics (QITL3)

2-4 June 2008
Helsinki, Finland

Linguistic Association of Finland (SKY)
and
Department of General Linguistics
University of Helsinki

Arppe, Antti; Kaius Sinnemäki; and Urpo Nikanne
(Editors)

Third Workshop on
Quantitative Investigations in Theoretical Linguistics **(QITL3)**
2-4 June, 2008
Department of General Linguistics, University of Helsinki, Finland
URL: `http://www.ling.helsinki.fi/sky/tapahtumat/qitl/`

Program Committee

Harald Baayen, University of Alberta
Marco Baroni, University of Trento/CIMeC
Peter Bosch, University of Osnabrück
Michael Cysouw, Max Planck Institute/Leipzig
Walter Daelemans, University of Antwerp
Stefan Evert, University of Osnabrück
Stefan Th. Gries, University of California, Santa Barbara
Stefan Grondelaers, Radboud University Nijmegen
Jennifer Hay, University of Canterbury
Timo Honkela, Helsinki University of Technology
Juhani Järvikivi, Max Planck Institute/Nijmegen
Brigitte Krenn, Austrian Research Institute for Artificial Intelligence (ÖFAI)
Jonas Kuhn, University of Potsdam
Merja Kytö, University of Uppsala
Roger Levy, University of California, San Diego
Anke Lüdeling, Humboldt University in Berlin
Elena Maslova, Bielefeld University
Detmar Meurers, Ohio State University
Matti Miestamo, University of Helsinki
Jussi Niemi, University of Joensuu
Martti Vainio, University of Helsinki
Yi Xu, University College London

Invited Speakers

Michael Cysouw, Max Planck Institute for Evolutionary Anthropology, Leipzig
Gary Marcus, New York University
Richard Sproat, University of Illinois at Urbana/Champaign

Organizing Committee

Laura Arola, University of Oulu
Antti Arppe, University of Helsinki, co-chair
Maria Metsä-Ketelä, University of Tampere
Maarit Niemelä, University of Oulu
Alexandre Nikolaev, University of Joensuu
Urpo Nikanne, Åbo Akademi University, co-chair
Kaius Sinnemäki, University of Helsinki, co-chair
Ulla Vanhatalo, University of Helsinki

Foreword and acknowledgements

The third Workshop on *Quantitative Investigations in Theoretical Linguistics* (QITL3), to be held on Monday-Wednesday, 2-4 June, 2008, in Helsinki, Finland, is co-hosted by the Linguistic Association of Finland (SKY) in association with the Department of General Linguistics at the University of Helsinki. This workshop is both a continuation of the two previous QITL events held in 2002 and 2006 in Osnabrück, Germany, and the latest in the sequence of summer symposia arranged annually by SKY.

We are grateful for a number of people and organizations for their support and assistance in making this Workshop happen. Among the previous organizers, Professor Anke Lüdeling, Humboldt-University of Berlin, and Professor Stefan Evert, University of Osnabrück, have provided us with invaluable encouragement and guidance in both getting the organization process started and at various stages along the way. Moreover, the support of Professor Fred Karlsson, University of Helsinki, for co-hosting this Workshop together with the Department of General Linguistics in many ways has facilitated its practical arrangement.

We were pleased to receive in all 36 submissions, though we could accept only one half (16) for oral presentation within the confines of our schedule, with a single general session according to QITL traditions. Consequently, we have incorporated for the first time also a Poster Session into the Workshop program for a further set of eight promising submissions. We look forward to hearing the final presentations and the kaleidoscope of linguistic research questions that they address with the help of quantitative methods, which is what this Workshop is essentially about.

With respect to long preparatory process leading to this Workshop, we want to thank each and every member of our Program Committee for their thorough reviews which guided us in evaluating the submitted abstracts and selecting from among them the ones for presentation, in addition to the constructively critical comments that were provided for all submissions, whether we could accept them into the program or not. We firmly believe that such feedback is crucial in building up the scientific quality of a workshop such as QITL3. Furthermore, we wish to thank our "local" members on the Program Committee, Juhani Järvikivi, Matti Miestamo, and Martti Vainio, for acting as a sounding-board for various decisions we took on the way concerning the makeup of the Workshop, though the final responsibility is naturally only ours. We also extend our thanks to Heidi Merimaa and Johanna Ratia for their assistance in practical administration and budgeting concerning this Workshop.

Finally, we also want to thank our governmental and corporate sponsors, the Academy of Finland, the University of Helsinki, Lingsoft, Connexor, and the Department of General Linguistics, who have each eased the financial burden in realizing QITL3.

Helsinki, May 2008

Antti Arppe	*Kaius Sinnemäki*	*Urpo Nikanne*
Laura Arola	*Maria Metsä-Ketelä*	*Maarit Niemelä*
Alexandre Nikolaev	*Ulla Vanhatalo*	

Table of Contents

Invited talks

Presentations

Invited Talks

Michael Cysouw
Max Planck Institute for Evolutionary Anthropology, Leipzig
cysouw@eva.mpg.de

Comparing the incomparable – Quantitative approaches for language comparison

When seriously looking at the world's linguistic diversity, languages are more different than often assumed. We might use the same names over and over to describe particular structures in different languages, but their forms and functions are always different. In this talk, I will start from the assumption that constructions are always language-particular, and thus that languages are in principle incomparable (a problem well-known in construction grammar).

Still, I will argue that it is possible to compare languages, when we accept that language comparison is crucially different from language- particular analysis. Specifically, language comparison might be 'besides the point' for individual languages, because the comparative view has to be kept constant. However, from this perspective, then, the internal organisation of individual languages will turn out to be the key to compare languages. Given a well-defined and invariable comparative perspective, the actual comparison of languages can be almost completely relegated to the summation over many individual language-particular analyses.

Gary Marcus
New York University
gary.marcus@nyu.edu

Language as *kluge*

In fields ranging from reasoning to linguistics, the idea of humans as perfect, rational, optimal creatures is making a comeback – but should it be? Hamlet's musings that the mind was "noble in reason ...infinite in faculty" have their counterparts in recent scholarly claims that the mind consists of an "accumulation of superlatively well-engineered designs" shaped by the process of natural selection (Tooby and Cosmides, 1995), and the 2006 suggestions of Bayesian cognitive scientists Chater, Tenenbaum and Yuille that "it seems increasingly plausible that human cognition may be explicable in rational probabilistic terms and that, in core domains, human cognition approaches an optimal level of performance", as well as in Chomsky's recent suggestions that language is close "to what some super-engineer would construct, given the conditions that the language faculty must satisfy".

In this talk, I will I argue that this resurgent enthusiasm for rationality (in cognition) and optimality (in language) is misplaced, and that the assumption that evolution tends creatures towards "superlative adaptation" ought to be considerably tempered by recognition of what Stephen Jay Gould called "remnants of history", or what I call evolutionary inertia. The thrust of my argument is that the mind in general, and language in particular, might be better seen as what engineers call a *kluge*: clumsy and inelegant, yet remarkably effective.

Richard Sproat
University of Indiana at Urbana/Champaign
rws@uiuc.edu

Experiments in morphological evolution

Morphology, inflectional morphology in particular, can often show surprising complexity in natural languages. In highly inflected languages, one often finds the situation where words fall into different inflectional classes, showing different markings for the same function. Syncretism – the same morphological marker serving several different functions – abounds, and sometimes this syncretism is systematic, motivating "rules of referral" that tie the expression of a particular set of morphosyntactic features to the expression of another. In a similar fashion, words may show seemingly arbitrary stem alternations, with particular stem variants being associated with particular slots in a paradigm: often these stem variants have no apparent phonological motivation.

How does such complexity come about?

I will attempt to provide partial answers by discussing some experiments in morphological evolution in multi-agent systems; this work follows very much in the tracks of previous work such as that of Kirby or Nettle or Wang and colleagues. One of the conclusions that seems to follow from this research is that some phenomena, such as "rules of referral", may be less interesting than their prominence in the linguistics literature might suggest. That is, the conditions that would motivate a linguist to posit a rule of referral can arise due to weak biases in the system, without any particular reference to global notions of the form "render slot X in the same way as slot Y".

Presentations

James P. Blevins, Farrell Ackerman, Paula Buttery, and Robert Malouf

University of Cambridge; University of California, San Diego;
University of Cambridge; San Diego State University

An entropy-based measure of morphological information

1. Introduction

Traditional approaches to morphology tend to treat inflectional systems not as unstructured sets of forms with shared stems or roots but as structured networks of elements. The interdependency of elements is, as Matthews (1991: 197) notes, 'the basis of exemplary paradigms' in the classical grammatical tradition. Although the exemplary patterns and leading forms of traditional descriptions bring out the structure of inflectional systems, traditional accounts are deficient – or at least incomplete – in a number of important respects. In particular, there is no method for measuring the implication structure of a set of forms or, no means of gauging the diagnostic value of specific forms within a set, and no generally accepted way even of identifying the leading forms of a system.

The approach outlined in this talk proceeds from the observation that implicational structure involves a type of **information**, specifically information that forms within a set convey about other forms in that set. Information in this sense corresponds to reduction in **uncertainty**. The more informative a given form is about a set of forms, the less uncertainty there is about the other forms in the set. In inflectionally complex languages, a speaker who has not encountered all of the forms of a given item is faced with some amount of uncertainty in determining the unencountered forms. If the choice of each form were completely independent, the problem of deducing unencountered forms would reduce to the problem of learning the lexicon of an isolating language. However, in nearly all inflectional systems, there are at least some forms of an item that reduce uncertainty about the other forms of the item. Once these notions are construed in terms of uncertainty reduction, the problem of measuring implicational structure and diagnostic value is susceptible to well-established techniques of analysis. The uncertainty associated with the realization of a paradigm cell correlates with its **entropy** (Shannon 1948) and the entropy of a paradigm is the sum of the entropies of its cells. The implicational relation between a paradigm cell and a set of cells is modelled by **conditional entropy**, the amount of uncertainty about the realization of the set that remains once the realization of the cell is known. The diagnostic value of a paradigm cell correlates with the **expected conditional entropy** of the cell, the average uncertainty remains in the other cells once the realization of the cell is known.

2. Information theoretic assumptions

In order to quantify the interrelations between forms in a paradigm, we will use the information theoretic notion **entropy** as the measure of predictability. This permits us to quantify "prediction" as a change in uncertainty, or information entropy (Shannon 1948). The idea behind information entropy is deceptively simple: Suppose we are given a random variable X which can take on one of a set of alternative values $x_1, x_2,\ldots,x_3$, with probability $P(x_1),P(x_2)\ x_2,\ldots,P(x_3)$. Then, the amount of uncertainty in X, or, alternatively, the degree of surprise we experience on learning the true value of X, is given by the entropy $H(X)$:

$$H(X) = -\sum_{x \in X} P(X) \log_2 P(X)$$

The entropy *H(X)* is the weighted average of the **surprisa**l $-\log_2 P(x_i)$ for each possible outcome *xi*. The surprisal is a measure of the amount of information expressed by a particular outcome, measured in bits, where 1 bit is the information in a choice between two equally probable outcomes. Outcomes which are less probable (and therefore less predictable) have higher surprisal. Specifically, surprisal is 0 bits for outcomes which always occur ($P(x) = 1$) and approaches ∞ for very unlikely events (as $P(x)$ approaches 0). The more choices there are in a given domain and the more evenly distributed the probability of each particular occurrence, the greater the uncertainty or surprise there is (on average) that a particular choice will be made among competitors and, hence, the greater the entropy. Conversely, choices with only a few possible outcomes or with one or two highly probable outcomes and lots of rare exceptions have a low entropy. One can also quantify the degree of prediction between cells using entropy. The average uncertainty in one variable given the value another is the **conditional entropy** *H(Y|X)*. If $P(y|x)$ is the conditional probability that $Y = y$ given that $X = x$, then the conditional entropy *H(Y|X) is:*

$$H(Y \mid X) = -\sum_{x \in X} P(X) \sum_{y \in Y} P(y \mid x) \log_2 P(y \mid x)$$

3. Implicational structure in Uralic

To demonstrate how an information-theoretic approach calculates the relative diagnosticity of words, the talk presents morphological patterns of ascending levels of complexity. The inflectional paradigms of Uralic languages are instructive because of the way that they realize inflectional properties by distinctive combinations of stem alternations and affixal exponence. Hence these systems are not amenable to a standard head-thorax-abdomen analysis in which lexical properties are expressed by the root, morphological class properties by stem formatives, and inflectional properties by inflectional affixes.

3.1 Northern Saami

First declension nouns in Northern Saami may inflect according to either of the patterns in Table 1.

Table 1: Gradation in first declension nouns in Saami (Bartens 1989: 511)

	'Weakening'		**'Strengthening'**	
	Sing	Plu	Sing	Plu
Nominative	**bihttá**	bihtát	baste	**basttet**
Gen/Acc	bihtá	bihtáid	**bastte**	**basttiid**
Illative	**bihttái**	bihtáide	bastii	**basttiide**
Locative	bihtás	bihtáin	**basttes**	**basttiin**
Comitative	bihtáin	bihtáiguin	**basttiin**	**basttiiguin**
Essive	**bihttán**		basten	
	'piece'		'spoon'	

In nouns of the 'weakening' type, the nominative and illative singular and the essive are all based on the strong stem of a noun, and the remaining forms are based on the weak stem. Nouns of the 'strengthening' variety exhibit a mirror-image pattern, in which the nominative and illative singular and essive are based on the weak stem,

and other forms are based on the strong stem. Strong forms, which are set in bold in Table 1, contain a geminate consonant which corresponds to a non-geminate in the corresponding weak forms. Given the paradigm in Table 1, we can calculate the conditional entropy of any one cell given any other cell. Take the nominative singular and the locative plural. Each has two possible realizations, and the entropy of each is 1 bit. To find the joint entropy, we look at the four possible combinations of realizations:

Nom Sg	Loc Pl	P
strong	strong	0.0
strong	weak	0.5
weak	strong	0.5
weak	weak	0.0

There are two equally likely outcomes, and the joint entropy is 1 bit. So the conditional entropy, *H(Loc Pl | Nom Sg)*, is 0 (*H(Nom Sg, Loc Pl) – H(Nom Sg)*).

That is, knowing the nominative singular realization for a particular lexeme completely determines the realization of the locative plural. We could repeat this calculation for any pair of cells in the paradigm and we would get the same result, as Saami nominal inflection is a completely symmetric system.

3.2 Finnish

The Finnish sub-paradigm in Table 2 illustrates a more typical pattern, in which different **combinations** of cells are diagnostic of declension class membership.

Table 2: Finnish *i*-stem and *e*-stem nouns (Buchholz 2004)

Nom Sg	Gen Sg	Part Sg	Part Pl	Ines Pl	
ovi	oven	ovea	ovia	ovissa	‘door’ (8)
kieli	kielen	kieltä	kieliä	kielissä	‘language’ (32)
vesi	veden	vettä	vesiä	vesissä	‘water’ (10)
lasi	lasin	lasia	laseja	laseissa	‘glass’ (4)
nalle	nallen	nallea	nalleja	nalleissa	‘teddy’ (9)
kirje	kirjeen	kirjettä	kirjeitä	kirjeissä	‘letter’ (78)

The implicational structure of the paradigms in Table 2 is set out in Table 3. The row expectation *E[row]* is the average conditional entropy of a column given a particular row. This is a measure of the **predictiveness** of a form. By this measure, the partitive singular is the most predictive form: if we know the partitive singular for a lexeme and want to produce another paradigm cell chosen at random, we will require only 0.250 bits of additional information on average.

Table 3: Conditional entropy *H(col | row)* of Finnish *i*-stem and *e*-stem nouns

	Nom Sg	Gen Sg	Part Sg	Part Pl	Ines Pl	*E[row]*
Nom Sg	—	1.333	1.667	0.874	0.541	1.104
Gen Sg	0.459	—	0.459	0.459	0.459	0.459
Part Sg	0.333	0.000	—	0.333	0.333	0.250
Part Pl	0.333	0.792	1.126	—	0.000	0.563
Ines Pl	0.459	1.252	1.585	0.459	—	0.939
E[col]	0.396	0.844	1.209	0.531	0.333	0.663

In contrast, given the nominative singular, we would need an additional 1.104 bit of information on average. The column expectation *E[col]* is the average uncertainty given a row remaining in a particular column. In contrast to the row expectations, this is a measure of the **predictedness** of a form. By this measure, the inessive plural is the most predicted form: if we want to produce the inessive plural for a lexeme and know some randomly selected other form, we will require on average another 0.333 bits of information.

Analyses of noun declensions in Tundra Nenets further confirm the value of entropy measures as a gauge of the implicational structure of a system and the 'diagnosticity' of individual elements. Entropy measures identify the leading forms of a system as the realizations that minimize the entropy of the system. The same measures also diagnose the anomaly of a fully suppletive class system or the pathologically 'uneconomical' classes of Carstairs (1983), since in neither type of system does knowledge about any form reduce the uncertainty of other forms. Furthermore, the application of standard information-theoretic techniques reinforces and helps to clarify previous implicative approaches to morphological analysis, such as Bochner (1993) and Finkel & Stump (2007). The use of entropy measures to analyze traditional notions like 'paradigm structure' (Wurzel 1970) also complements the use of these measures to model response latencies in psycholinguistic research (Moscoso et al. 2004, Milin et al. to appear).

References

Bartens, H.-H. (1989). *Lehrbuch der saamischen (lappischen) Sprache*. Helmut Buske Verlag.

Bochner, H. (1993). *Simplicity in generative grammar*. Mouton de Gruyter.

Carstairs, A. (1983). Paradigm economy. *Journal of Linguistics* **19**, 115–125.

Finkel, R. & Stump, G. (2007). Principal parts and linguistic typology. *Morphology* **17**, 39–75.

Buchholz, E. (2004). *Grammatik der finnischen Sprache*. Bremen: Hempen Verlag.

Matthews, P. H. (1991). *Morphology*. Cambridge University Press.

Milin, Petar, Filipović, Đurdević & Moscoso del Prado Martín, Fermín (to appear). The simultaneous effects of inflectional paradigms and classes on lexical recognition: Evidence from Serbian. *Journal of Memory and Language*.

Moscoso del Prado Martín, Fermín, Kostić, Aleksandar, &Baayen, R. Harald (2004). Putting the bits together: an information-theoretical perspective on morphological processing. *Cognition* **94**, 1–18.

Shannon, C. (1948). A mathematical theory of communication. *The Bell System Technical Journal* **27**, 379–423, 623–656.

Wurzel, W.U. (1970) *Studien zur deutschen Lautstruktur*. Berlin: Akademie-Verlag.

Dagmar Divjak
Science Foundation – Flanders (Belgium) & University of Sheffield (UK)
d.divjak@sheffield.ac.uk

Modelling aspectual choice in Polish modal constructions.

Introduction

Much effort has been put into clarifying the relation between modality and other verbal properties, in particular mood and tense. Until recently the relation between modality and aspect received much less attention, however. For Slavic languages this situation is particularly unfortunate as Slavic languages mark the imperfective versus perfective distinction on all verbal forms, so there is no avoiding aspectual choice. Moreover, the hypothesis that directed much of the research, i.e. that imperfective aspect prevails in modal constructions or that the imperfective is used to express epistemic or alethic modality whereas perfective aspect renders deontic meanings, does not seem to hold for Slavic data: it has long been recognized that, if anything, the perfective would used more frequently in modal constructions in general (cf. Rassudova 1968, Forsyth 1970) and the imperfective aspect would be preferred in deontic contexts (Padučeva 2006, Šmelev & Zalizniak 2006, Wiemer ms.).

An exploratory comparative study of Russian, Polish and Serbian

An exploratory cognitive linguistic, corpus-based, quantitative study was carried out to identify the aspectual preferences of dynamic (participant inherent vs participant imposed) and deontic modality (Nuyts 2006) in positive and negative declarative sentences of the type exemplified in (1) and (2).

(Russian)

(1) *Zdes'* *možno* *perechodit' ulicu*
Here.$_{ADV}$ possible/permissible.$_{PREDADV}$ cross.$_{IMPF.INF}$ street.$_{ACC.F.SG}$
'You can cross [permissibility] the street here'

(2) *Zdes'* *možno* *perejti ulicu*
Here.$_{ADV}$ possible/permissible.$_{PREDADV}$ cross.$_{IMPF.INF}$ street.$_{ACC.F.SG}$
'You can cross [possibility] the street here'

Starting point for the comparative study was the situation in Russian that provides *možno/nel'zja* to express (in-)ability, (im)possibility and (non-)permissibility and *nužno* and *nado* to express necessity and obligation. On the basis of data extracted from a 1 million word parallel Slavic corpus compiled specifically for this study (see Table 1), Polish and Serbian translational equivalents were identified (12 for Polish, 7 for Serbian) to facilitate a direct comparison with the findings for Russian. In all, the 983 retrieved instances are tagged for language, novel, author/translator, modal word, aspectual range of the infinitive (impf only, pf only, biaspectual, impf_pf), aspect of the infinitive (impf vs pf), modality type (dynamic vs deontic) and polarity (positive vs negative).

Table 1. Corpus contents

Original	Translation	Translation
(Russian) Bulgakov, M. 1938. *Master i Margarita.*	(Polish) Mistrz i Małgorzata (by Irena Lewandowska & Witold Dąbrowski)	(Serbian) Мајстор и Маргарита (by Milan Čopić)
(Polish) Lem, S. 1961. *Solaris*	(Russian) Солярис (by Dmitrij Bruškin)	(Serbian) Solaris (by Predrag Obućina)
(Serbian) Pavić, M. 1984. *Hazarskij Rečnik.*	(Polish) Słownik chazarski (by Elżbieta Kwaśniewska & Danuta Cirlić-Straszyńska)	(Russian) Хазарский словарь (by Larisa Savel'evaja)

Given the make-up of the corpus, the observations cannot be considered independent, hence mixed effects modelling (with Novel and Modal Word as random effects) using lmer (Baayen 2008: ch. 7) was carried out on the 830 instances that contain an infinitive that exists in both imperfective and perfective, i.e. allow aspectual choice. The results of the best performing model are summarized in Table (2).

A model with language and modal word as random effects and modality type plus polarity as fixed effects revealed that, in all three Slavic languages studied, 1) in general, perfective infinitives were used significantly more frequently in modal declarative sentences built around a modal word followed by an infinitive than imperfective infinitives; 2) it is significantly less likely to find a modal adverb followed by an perfective infinitive when deontic modality is expressed than it is to find an imperfective infinitive; 3) it is significantly more likely to find a perfective infinitive when the modal statement is positive than it is to find an imperfective infinitive.

Table 2. Comparing models across languages

Russian	Polish	Serbian
a modal adverb followed by a perfective infinitive is used to express deontic modality [estimate = -5.4567, p= 6.95e-11]	a modal adverb followed by a perfective infinitive is used to express deontic modality [estimate = -2.1838, p= 1.5e-06]	a modal adverb followed by a perfective infinitive is used to express deontic modality [estimate = -2.8217, p= 3.53e-09]
a modal adverb followed by a perfective infinitive is found when the modal statement is positive [estimate = 3.8689, p= 0.000807]	a modal adverb followed by a perfective infinitive is found when the modal statement is positive [estimate = 0.7439, p= 0.05308]	a modal adverb followed by a perfective infinitive is found when the modal statement is positive [estimate = 1.3420, p=0.000362]
Estimated scale [0.9864484]	Estimated scale [0.991989]	Estimated scale [0.980616]
C index of concordance [0.8670398]	C index of concordance [0.7405442]	C index of concordance [0.8037842]
Somer's D [0.7340796]	Somer's D [0.4810883]	Somer's D [0.6075684]

Although both modality and polarity show up as significant predictors of the choice of a particular aspect for the infinitive in all three languages, the model fits Russian best. This outcome is expected on Dickey's (2000) division of the Slavic aspectual world: with the Russian aspectual system focused on definiteness in time, the imperfective expresses "qualitative temporal indefiniteness", i.e. lack of assignability to a single, unique point in time, which fits well with the "general timeless applicability" of deontic modality. Polish and Serbian being transitional zones between the Eastern and Western systems, they likewise display the pattern observed for Russian, albeit to a lesser extent. Polish, although predicted to be more similar to Russian than Serbian, seems to deviate in particular from the expected aspectual pattern: Somer's D reveals

only a medium rank correlation between predicted probabilities and observed responses while the obtained C index of concordance remains below the 0.8 threshold, generally required to recognize the predictive power of a model; this performance is particularly poor given that the percentage of correctly predicted cases would be about 75%, merely by selecting perfective infinitives in all cases, and not including any predictor variables (Johnson 2008: 254-255). Yet, regarding the perfective as the "default" aspect for modal contexts would be highly unusual: in Slavic languages, perfective aspect is the marked member of the opposition, and marked members would typically be expected to occur less frequently and in fewer contexts than their unmarked counterpart (Forsyth 1970: 6-8).

A model for Polish – and other Slavic languages?

In order to arrive at an adequate model of aspect assignment in modal constructions in Polish, the corpus sample used was increased (from 240 to 400 examples) while at the same time the number of modal predicative adverbs was decreased (from 12 to 7). Moreover, 4 additional properties were taken into account. These properties relate to the semantics of the modal word (the modality type expressed, i.e., possibility vs permissibility vs necessity vs obligation vs ability vs volition vs prediction), and of the aspect of the infinitive (the aspectual type rendered, i.e., generalizing vs specifying use and activity focused vs result focused) as well as to the degree of control (high, medium, low) the subject has over the infinitive action.

A new mixed effects logistic regression model (again with Novel and Modal Word as random effects) was fit to the corpus data in order to reveal the variable or set of variables that has the highest predictive power for aspect assignment in modal constructions. The results of the best performing model are summarized in Table (3).

Table 3. A new model for Polish.

Polish (old)	Polish (new)
it is significantly less likely to find a modal adverb followed by an perfective infinitive when deontic modality is expressed [estimate = -2.1838, p= 1.5e-06]	it is significantly more likely to find a modal adverb followed by a perfective infinitive when dynamic modality is expressed [estimate = 1.0474, p= 0.00955]
it is marginally significantly more likely to find a perfective infinitive when the modal statement is positive [estimate = 0.7439, p= 0.05308]	it is significantly more likely to find a modal adverb followed by an imperfective infinitive when a generalization is expressed [estimate = 3.6962, p= < 2e-16]
Estimated scale [0.991989]	Estimated scale [0.9748724]
C index of concordance [0.7405442]	C index of concordance [0.9016152]
Somer's D [0.4810883]	Somer's D [0.8032304]

Although type of modality remains a significant contributor to aspectual choice, the fact whether the option, permission, order etc. has been given to carry out an action only once (aka specifying use) or multiple times (aka generalizing use) outperforms the type of modality in predicting the choice of aspect for the infinitive.

Theoretical implications

This study revealed that quantitative corpus-linguistic methodologies capable of honoring the multifaceted nature of the phenomenon under investigation might

necessitate rejecting theoretically motivated models in favor of cognitively simple(r) models. The initial outcome suggested that the "lexical" meaning of modality (dynamic vs deontic) as well as polarity (positive vs negative) predict aspectual choice in modal constructions quite well, at least for Russian and to a lesser extent Serbian. This finding reverses the claims made in the general linguistic literature while confirming the corrections proposed by Slavic linguists. Yet, an in-depth study of Polish revealed that other variables might be better at predicting aspectual choice: the "grammatical" meaning of aspect as captured by the parameter specific vs generalizing outperforms the "lexical" meaning of modality when it comes to predicting aspectual choice, and makes polarity superfluous.

On a cognitive linguistic approach, this outcome comes as no surprise. A cognitive approach to aspect assumes that the semantics of aspectual categories is organized around a prototype with many language-particular extensions, including extensions in other domains such as tense and modality. In this case, the "grammatical" meaning of aspect extends flawlessly into the "lexical" meaning of modality. Dynamic modality is concerned with a particular situation or a participant in that situation, hence quite similar to the prototypical interpretation of perfectively coded events as having summarizing properties and as presenting situations as one-off events or as events with specific settings. Deontic modality, on the other hand, regulates existence for everyone, always and everywhere, hence expresses a meaning that is similar to the prototypical interpretation of the imperfective as encoding statements of fact, as events with focus on the process or as repeated events. Further research will show whether the same relation between aspect and modality holds in other Slavic languages and theoretical models of aspect/modality interaction should be adapted accordingly.

References

Baayen, R.H. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics using R.* Cambridge: Cambridge University Press.

Dickey, S. 2000. *Parameters of Slavic Aspect: A Cognitive Approach*. Stanford: Center for the Study of Language and Information.

Forsyth, J. 1970. *A Grammar of Aspect. Usage and Meaning in the Russian Verb.* Cambridge: Cambridge University Press.

Johnson, K. 2008. *Quantitative Methods in Linguistics*. Oxford: Blackwell Publishing

Nuyts, J. 2006. Modality: Overview and linguistic issues. In: *The expression of modality*, W. Frawley (ed.), 1-26. Berlin: Mouton.

Padučeva, E. 2006. *Modality, Negation and Aspect: the case of the Russian* možet *and* dolžen*.* Paper presented at the 39th Annual Meeting of the SLE, Bremen.

Rassudova, O. P. 1968. *Upotreblenie vidov glagola v russkom jazyke.* Moskva: Izdatel'stvo Moskovskogo Universiteta.

Šmelev, A. & Zaliznjak, A. 2006. *Aspect, Modality and closely related categories in Russian.* Paper presented at the Inaugural Meeting of the Slavic Linguistic Society in Bloomington, IN.

Wiemer, B. (ms.). *Aspect choice in modal and pragmatic contexts: pieces of a puzzle in Russian and other Slavic languages.*

Nigel Duffield[1], Ayumi Matsuo[1] and Leah Roberts[2]
University of Sheffield[1] & Max Planck Institute for Psycholinguistics[2]
Leah.Roberts@mpi.nl[2]

Seeing what's missing: What (eye-tracking) data from native speakers and second language learners can tell us about the theoretical distinction between VP-ellipsis and VP-anaphora

VP-ellipsis is an interface phenomenon *par excellence*: the constraints on the grammatical acceptability of ellipsis clauses are at once purely syntactic—grammaticality is determined by the *structural* properties of the antecedent clause—and at the same time discourse-dependent—in contrast to some other core syntactic phenomena, the grammatical acceptability of an ellipsis clause cannot be determined without reference to the preceding linguistic discourse. At least since Sag's seminal work on the topic Sag (1976), VP-ellipsis facts have been central to the development of grammatical theory, especially in the context of Minimalist concerns with interface conditions, and this theoretical work continues to progress our understanding of the limits of core grammar; see, for example, Hankamer & Sag (1976), Sag & Hankamer (1984), Lobeck (1995), Johnson (1997), Merchant (2001), *inter alia*.

In tandem with purely theoretical work, experimentalists in adult language processing and first and second language acquisition have attempted to flesh out our understanding of the effects of constraints on VP-ellipsis in language processing, and to determine how and when these constraints are acquired by different groups of learners. One particular property that has received psycholinguistic attention is the so-called Parallelism Constraint (Hankamer & Sag 1984), the requirement that the ellipsis clause be syntactically parallel to its antecedent, as in (1a) vs. (??1b); a constraint, which—at least according to the theoretical literature—does not apply to (semantically-equivalent) VP-Anaphora constructions (2ab):

1. a. Someone had to put out the garbage, but John didn't want to.
   b. ??The garbage had to be put out, but John didn't want to.
2. a. Someone had to put out the garbage, but John didn't want to do it.
   b. The garbage had to be put out, but John didn't want to do it.

Tanenhaus & Carlson (1990) report a set of studies using the Sentence Completion Judgment Paradigm—a timed reading task in which subjects are asked to decide whether the second sentence (ellipsis clause) follows naturally and meaningfully from the first (Antecedent clause). These results show that adult native speakers are indeed sensitive to the effects of syntactic parallelism in language processing, both in the case of active vs. passive antecedents, as in (1)/(2), *as well* in the case of verbal vs. nominal antecedents (in 3 and 4 below).

3. a. John wanted someone to kiss him, but Jo didn't want to.
   b. ?*John wanted a kiss, but Jo didn't want to.
4. a. John wanted someone to kiss him, but Jo didn't want to do it.
   b. John wanted a kiss, but Jo didn't want to do it.

In a series of papers, Duffield & Matsuo further develop this paradigm, extending it to investigate Second Language learners' knowledge of ellipsis constraints—see Duffield & Matsuo (2000, 2001, 2002 ,Duffield & Matsuo (2003). The results of these latter experiments reveal two new findings: first, that lay native-speakers' acceptability judgments differ systematically from those given in most theoretical work—for example, there *is* a parallelism effect for VP-anaphora constructions also (albeit a smaller one); second, that advanced Dutch L2 learners, whose L1 does not license VP-ellipsis, *can* acquire the relevant constraints, but that their judgments nevertheless differ systematically from those of native-speakers when more specific factors (finiteness, construction type, semantic recoverability) are analyzed in detail.

A potential criticism of these results, however, is that Sentence Completion Judgment is not a true 'online' task: since it only measures responses at the offset of the stimulus sentence, it does not directly tap the use of grammatical knowledge in online processing. To address this, we employed the same materials used in Duffield & Matsuo's SCJ experiments in a reading task using a head-mounted eye-tracker (cf. Tanenhaus et al. 2000). Eye-tracking technology affords a number of different dependent measures that together provide a millisecond-by-millisecond chart of a subject's reading of a given stimulus sentence: *first fixations and first-pass reading times*, which are thought to reflect the earliest stages of processing, and which measure the amount of time each subject initially fixates on critical positions in the sentence; *regressions*, which show subjects' returns to earlier leftward positions—for example, to a potential antecedent phrase; *second pass fixations and total reading times*, which together reflect later stages of processing and integration.

Using this technology, we tested 15 English native-speakers and 20 advanced Dutch L2 learners of English on Duffield & Matsuo's materials, manipulating (in the test sentences) the syntactic parallelism of the antecedent clause (active/*passive/*nominal) and the anaphor type of the second sentence (VPE/VPA). Given the previous results from SCJ tasks, we predicted that English native-speakers should show reliable effects of parallelism in the VPE condition for both the active/passive and the verbal/nominal antecedent types, but expected the effect to be stronger for nominal vs. passive antecedents. For the Dutch L2 learners, since VPE is ungrammatical in their L1, and if they have yet to acquire the required competence in English, then we expected no such asymmetry in the parallelism effect, that is, both ellipsis types should be equally difficult to process. For both early measures, the analysis found a main effect of Ellipsis Type and no interaction with the between-subjects factor Language Group (First fixation durations: F_1 (1, 33) = 5.52; $p < 0.03$; $\eta^2 = .14$; F_2 (1, 11) = 6.85; $p < 0.03$; $\eta^2 = .38$; First pass times: F_1 (1, 33) = 4.88; $p < 0.04$; $\eta^2 = .13$; F_2 (1, 11) = 17.02; $p < 0.003$; $\eta^2 = .61$): irrespective of the type of antecedent, both the native English speakers and the L2 learners spent more time fixating the critical region in the VPE constructions (First Fixations: 236 ms; First Pass Times: 262) than in the VPA constructions (First Fixations: 223 ms; First Pass Times: 238). The expected interaction between Anaphor Type and Ellipsis type showed up in the later measures (Second Pass Fixations: F_1 (3, 99) = 3.71; $p < 0.03$; $\eta^2 = .26$; F_2 (3, 33) = 4.11; $p < .02$; $\eta^2 = .27$; Total Fixation Durations: F_1 (3, 99) = 7.37; $p < 0.001$; $\eta^2 = .42$; F_2 (3, 33) = 3.64; $p < .03$; $\eta^2 = .25$); here, again there was no difference between the two subject groups.

In summary, these new results provide confirmation of the claim that the parser has early and continuous access to grammatical information about constraints on VP-ellipsis constructions, and that such knowledge is both acquirable—and *used*—by advanced L2 learners online in L2 processing (subtle differences in the use of this knowledge notwithstanding).

Selected References

DUFFIELD, NIGEL and MATSUO, AYUMI. 2002. Finiteness and Parallelism: assessing the generality of knowledge about English ellipsis in SLA. Proceedings of the 26th Boston University Conference on Language Development, ed. by Barbora Skarabela, Sarah Fish and Anna H.-J. Do, 197-207. Somerville, MA: Cascadilla Press.

—. 2003. Factoring out the parallelism effect in VP-ellipsis. Proceedings of the 39th Regional Meeting of the Chicago Linguistic Society: Main Session, ed. by Jonathon Cihlar. Chicago: Chicago Linguistic Society.

HANKAMER, JORGE and SAG, IVAN. 1976. Deep and surface anaphora. Linguistic Inquiry, 7.391-428.

JOHNSON, KYLE. 1997. When verb phrases go missing, ms. University of Massachusetts, Amherst.

LOBECK, ANNE. 1995. Ellipsis. Oxford: Oxford University Press.

MERCHANT, JASON. 2001. The syntax of silence: Sluicing, islands, and the theory of ellipsis. Oxford: Oxford University Press.

SAG, IVAN. 1976. Deletion and logical form, MIT: doctoral dissertation.

SAG, IVAN and HANKAMER, JORGE. 1984. Towards a theory of anaphoric processing. Linguistics and Philosophy, 7.325-45.

TANENHAUS, MICHAEL and CARLSON, GREG N. 1990. Comprehension of deep and surface verbphrase anaphors. Language and Cognitive Processes, 5.257-80.

Annarita Felici and Paul Pal
Royal Holloway, University of London
A.Felici@rhul.ac.uk, P.Pal@rhul.ac.uk

A probabilistic approach to language structure

1. Introduction

The translation of international legal instruments requires a high degree of accuracy and consistency. With the increasing demand for multilingual texts, translation memory tools and research on parallel corpora have proved to be particularly useful for the translation of repetitive documents, as well as for those subject to an evolutive drafting process and production. Moving from this assumption, the present study adopts a probabilistic approach to the comparison of some repetitive language structures in multilingual legal texts. Data (1.404.723 words) consists of a multilingual parallel corpus in four languages: English, French, German and Italian. All the documents have been taken from the EU secondary legislation and include *Regulations, Decisions, Directives* and *Recommendations*, chosen between the years 2001-04. Texts are all strictly 'normative' and discourse is expected to be precise with minimum scope for ambiguity. The main focus is prescriptive statements, namely *deontic* norms (permission, obligation, prohibition) and constitutive performatives. Their formulation is highly standardized in English, both within and outside the EU context (Coode 1843, Driedger 1976, The EU Interinstitutional Style Guide), and modal verbs play a consolidated pivotal role. On the other hand, their expression in other languages is more vague and extensive, with potential consequences on the translation of norms. Bearing these remarks in mind, our objective is: 1) to evaluate the degree of prescriptive standardization with reference to English and the other three languages, and 2) to predict the general pattern of expression in the other languages under the condition that (i) English legal drafting is highly standardized, (ii) the EU and the main English drafting guidelines tend to use modal verbs in prescriptive statements (iii) text types under examination are repetitive and reusable (iv) the four EU instruments can be more or less binding. English is used as the main entry point and entropy analysis is exploited to measure the number of alternatives (degree of uncertainty) occurring in the other three languages. By adding knowledge to a system (e.g. a more standardized formulation), one reduces the number of alternatives (uncertainty), which leads to a decrease of entropy and to a gain of information in the expression of the norm. Although language phenomena cannot be fully described, the results of this analysis have empirically proved that given a set of conditions, certain linguistic structures are more easily predictable than other when comparing several languages. These types of analysis can foster research in language testing, evaluation, and in the development of automated translation's tools.

2. Theoretical background and probabilistic variables

Normative sentences can take different grammatical and lexical forms. The main verb usually determines the type of norm that is to be expressed (e.g. obligation, permission, empowering, prohibition) and following Austin (1962) can be 'explicit' (order, permit,

forbid) or 'implicit' (*shall, may, must*). If the same information is communicated through different languages, the property of one individual language, including its structure, might be reflected in the text. However, this is not always the case. Different languages have different modes of expressions and drafting conventions may impact stronger than grammar on legal discourse. For the purpose of this specific analysis, we chose a EU *Regulation* sub-corpus in the 4 languages (334.425 words in total)[1]. With the help of Paraconc we initially retrieved all the English modal verbs inherent to the expression of norms, together with their translation equivalents in French, German and Italian. Corpus findings confirmed the predominant occurrence of the English modals *shall, must, may* and to a lesser extent *can* and *should.* The other three languages showed a variety of linguistic forms (alternatives) that are grouped as follows: (a) present indicative, (b) modal verbs, (c) verbal periphrases, (d) lexicalized modal expressions, (e) ellipsis or zero correspondence. In order to apply probabilistic treatment, we selected 5 categories of expressions corresponding to each modal verb or language alternative. They include: (a) constitutive norms and obligations, (b) logical necessity, (c) permission and authorization, (d) capability and (e) non-binding norms. The probabilistic approach starts with determining the frequency of occurrence (n_i say) of each linguistic form (modals and other linguistic alternatives) associated with a category. A probability variable p_i is then derived from the estimated proportion of occurrence of a particular modal verb in the corpus. This is given by $p_i = n_i / n$, where n is the total number of modals or their equivalents. Referring to the English *Regulation*, the five probabilities are expressed as follows:

$p_1 = p_{mv \rightarrow shall} = n_{shall} / n$; $p_2 = p_{mv \rightarrow must} = n_{must} / n$; etc.

In French, German and Italian texts of the same document they are expressed as:

$p_1 = p_{pres.ind.} + p_{mv} + p_{vp} + p_{me} + p_{ellipses}$; $p_2 = p_{pres.ind.} + p_{mv} + p_{vp} + p_{me} + p_{ellipses}$ etc.

The modal *shall* is the most frequent auxiliary to impose obligations and binding norms while *may* is used to express permissions and authorizations. From a statistical point of view, variations in the linguistic forms of expression are possible due to the number of alternatives inherent in a language.

In the information theory, the metric used to measure information is known as *entropy* (h) and corresponds to a degree of uncertainty (a shortage of information due to the large numbers of alternatives) in a message. According to Shannon (1949), the information value or content *h(p)* is dependent on the probability of occurrence (*p*) of an event. This dependence is described by the formula: $h(p) = - \log (p) = \log (1/p)$. Different languages in their repertoire have different linguistic forms, and therefore each mode carries different probability values. The more precise or standardized the system is, the less its entropy (e.g. the number of alternatives) is. Considering the EU *Regulation* document, the probability p_i of occurrence of each individual form (e.g. pres. ind, mv, vp, me and ellipses) belonging to the 5 categories of norms is linked to a certain information value. The sum of these probabilities over all the distinct forms produces different results and hence different information values. The expected

[1] The whole study includes the four EU secondary legislation text types. For reason of space, we are presenting only data concerning the Regulation that is the most binding text out of the four.

information content of a system is the sum of the information contents weighted by the probabilities of the respective constituent attributes. This sum is expressed as follows:

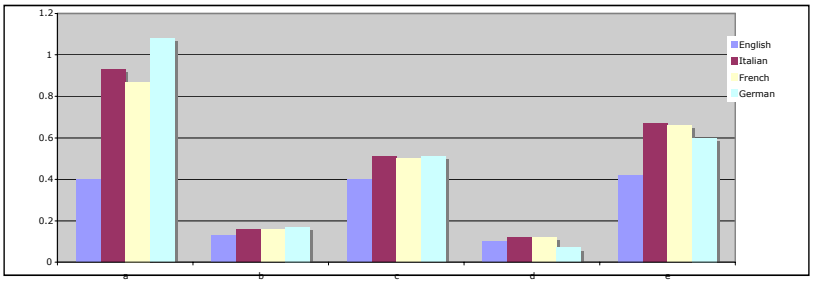
$$H = -\left[\sum_{i=1}^{i=5} p_i \log_2(p_i)\right]$$

where p_i is the probability belonging to a certain category of expression *i* (e.g. (a) constitutive norms and obligations, (b) logical necessity, (c) permission and authorization etc).

3. Entropy results

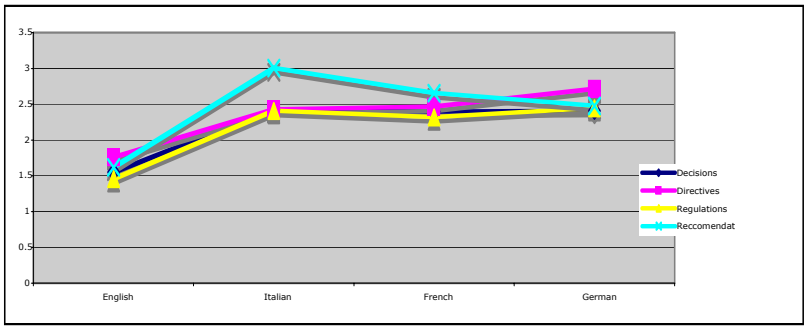
Entropy evaluation has been carried out at two levels: (1) entropy measures with respect to the 5 categories of expression (constitutive and performative norms, logical necessity, permission and authorization, capability and non-binding norms) as in figure 1; and (2) overall entropy in the EU *Regulation* and in the whole secondary legislation corpus as showed in figure 2.

Figure 1. Entropy measures in the EU Regulation (5 category of expressions)



The different heights for French, German and Italian reflect in each cluster the use of alternatives in these languages compared to the English system, where only modal verbs have been considered.
In the material examined, the overall entropy of a language is the sum of the separate entropy measures with respect to the modal verbs as found in the different linguistic versions of the corpus. Extending this approach to norm formulation in the 4 types of EU documents (*Decisions, Directives, Regulations* and *Recommendations*), we were able to determine the overall entropy for each text type in the 4 languages and compare their linguistic alternatives.

Figure 2. Overall entropy in the EU secondary legislation corpus



In this case entropy results provide measures of particular EU text types and can confirm inference on their degree of mandatory force.

4. Discussion

By applying entropy analysis, four language systems have been compared in the attempt to ascertain the degree of prescriptive standardization occurring in a relative small corpus of EU normative texts. English modal verbs serve here as a parameter for their consolidated position in the international legal drafting and also for English being the main working language of the EU. Its entropy results are therefore lower when compared to the other three languages and do not constitute a relevant asset to the goal of this analysis. From figure 1, it is possible to remark that the formulation of logical necessity, permissions and authorization and capability is quite standardized in the four languages. For each English modal, it is to be expected an equivalent modal verb in French, German and Italian. This is not the case of the constitutive and performative norms where a hypothetical translator or translation tool is exposed to a considerable variation. The three languages account for a larger number of alternatives against the English *shall*, with an overall preference for the present indicative. This is partly due to the widespread use of *shall* in the English EU drafting, but also to the inherent complexity of these norm types, which can indicate definitions, constitutive performatives, obligations and prohibitions. It is also interesting to remark that although French and Italian boast a similar semantic and grammatical language system, entropy results are not as close as in the other categories. This is probably due to the more prominent role of French in the EU context and, hence to an increased standardization. The lower entropy results in the German non-binding norms are due instead to the established position of the conditional form of the modal *sollen* when formulating general guidelines and recommendations. Figure 2 gives entropy results on the basis of the four EU legal instruments and text types. *Regulations* and *Decisions* present in the four languages lower entropy because the direct applicability of norms requires more precision and a more standardized formulation. Again, French *Regulations* and *Decisions* account for slightly minor entropy than Italian and German. On the other hand, the *Recommendations* text type highlights several alternatives above all in French and Italian, whose figures look closer in these respects. In conclusion, the application of probabilistic theories has proved that given certain conditions, it is possible to predict with some degree of certainty the occurrence of a particular factor. When applied to parallel texts, entropy analysis can delve into theoretical issues about language structure, but can also provide a resourceful ground of applications for language testing and evaluation of machine translations and other automated translation tools.

References

Austin, John Langshaw 1962. *How to do things with words.*Oxford: Oxford University Press.

Coode, George.1843. *Legislative Expressions. Appendix to the Report of the Poor Law Commissioners on Local Taxation*. Published separately 1845, 2nd Ed. 1852.

Driedger, Elmer A. 1976. *The Composition of legislation.* Legislative forms and precedents (2nd Ed.). Ottawa:The Department of Justice

Shannon, Claude and Weaver Warren.1963 (1949) *The mathematical theory of communication.* Urbana: University of Illinois Press.USA.

http://publications.europa.eu/code/en/en-6000000.htm (accessed on 27.01.2008)

Dylan Glynn
University of Leuven
dylan.glynn@arts.kuleuven.be

Clusters and Correspondences. A comparison of two exploratory statistical techniques for semantic description

Introduction

Corpus-Driven quantitative techniques for language description have witnessed important success in recent years. In semantic research, the main of this drive has been in disambiguation, whether on the syntagmatic or paradigmatic plane. Many have now taken the next step, seeking to employ such techniques for the description of lexical semantic structure *per se*. This study examines two exploratory multivariate statistical techniques, namely Multiple Correspondence Analysis (MCA) and Hierarchical Cluster Analysis (HCA), and considers the strengths and weaknesses of both approaches for the description of lexical semantic variation.

This study is informed by the usage-based approach of Cognitive Linguistics, represented by Geeraerts & al. (1994), Gries (2003), Geeraerts (2006), Tummers & al. (2005), Gries & Stefanowitsch (2006), Grondelaers & al. (2007), and Heylen & al. (in press). Within this field, both exploratory and confirmatory statistical techniques enjoy wide currency. Four of the main techniques include Cluster Analysis (Divjak 2006, Divjak & Gries 2006, Gries 2006), Correspondence Analysis (Arppe 2006, Glynn in press, Gries & David forthc.) for exploratory research and Logistic Regression Analysis (Heylen 2005, Tummers & al. 2005) and Linear Discriminant Analysis (Gries 2001, Wullf 2003) for hypothesis testing.

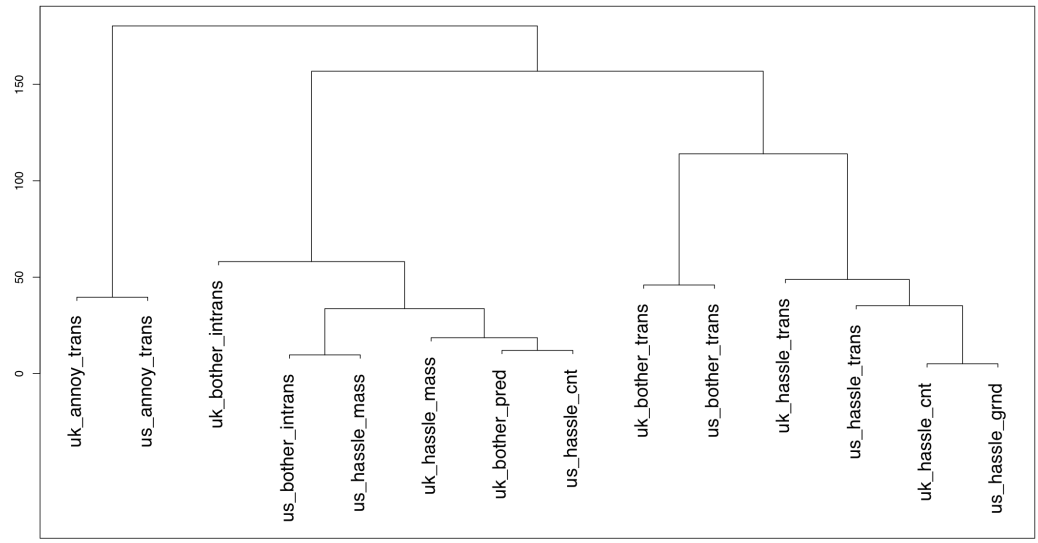
Case Study

Using a large non-commercial corpus, built from on-line personal diaries and specified for the regional difference (American vs. British English), the study examines the semantic structure of the lexeme *annoy*, semasiologically and onomasiologically in comparison with two parasynonymous words; *hassle* and *bother*. Approximately 500 occurrences are manually annotated for 20 formal, semantic, and extra-linguistic variables. An important challenge for corpus linguistics is semantic description. In order to maximise objectivity of the semantic annotation, special attention is given to the Frame Semantic actor types and their relations. This method has been shown to provide indirect indices of semantic structure (Glynn & al. forthc). Despite the regional variation, the corpus is quite homogenous in terms of register and theme. However, this allows us to focus on the dimension of dialect variation, relative to the variables of morpho-syntax and Frame Semantic argument structure.

At an exploratory level, both MCA and HCA have important strengths and weaknesses. One important difference between the two techniques is that Cluster Analysis is primarily designed to present its results in the form of dendograms where Correspondence Analysis relies on scatter plots. The dendograms of HCA offer clear representations of both the grouping of features and the relative degree of correlation between those features. The trees represent relations and the shorter the distance between the node and the branch, the higher the degree of correlation. The principle shortcoming of this representation it that gives the false impression that all the data fall into groups, where in fact this may not be the case. Figure 1, offers an example of the

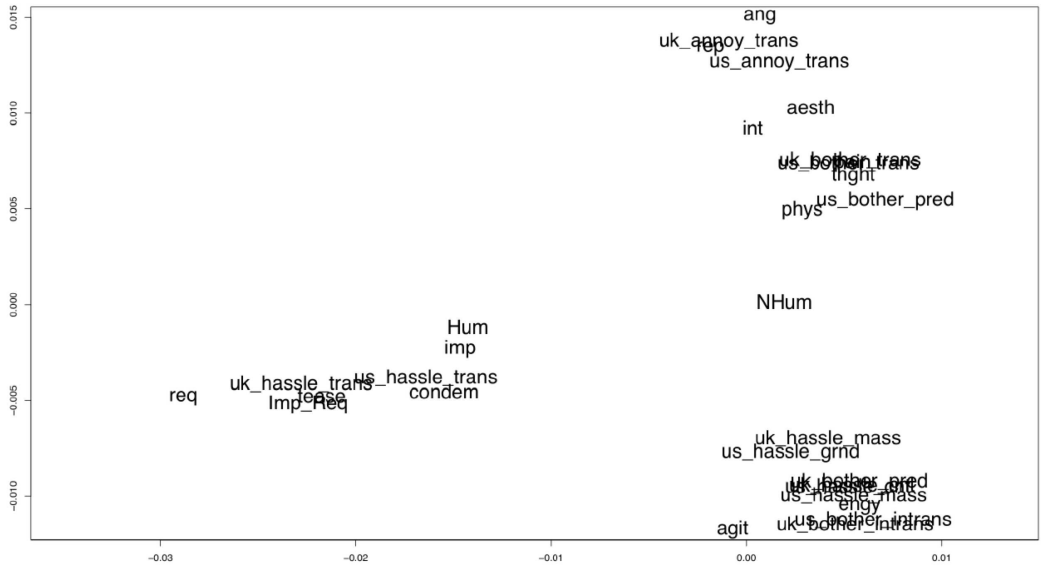
dendogram of HCA. The table clusters the lexmes – *annoy* – *bother* - *hassle*, with by variables of dialect – grammatical construction, grammatical class, cause and affect of the event.

Figure 1. HCA *annoy-hassle-bother*



The scatter plots of Correspondence Analysis, although at times difficult to interpret, offer a much more 'analogue' representation of correlation. In such plots, the spatial dispersion and relative proximity of data points represent degrees of correlation. The interpretation of the plot is then much more approximative than the dendogram. Figure 2. offers an example of the visurlaisation of an MCA analysis for the same variabels that are treated above.

Figure 2. MCA *annoy-hassle-bother*



The results between the two techniques here seem somewhat in contrast. In order to better appreciate the differences. A simpler dataset is analysed. Looking at just *hassle*, we get a more comparable result. Agent and Patient Types, Agent and Patient Person (1st, 2nd, 3rd) as well as Agent-Patient relations differ significantly in the use of *annoy* between the two dialects. Moreover, these differences are mirrored by purely semantic variables. The combination of the Frame Semantic, formal and traditional semantic variables show that *annoy* possesses a more emphatic and 'anger' related meaning in American than in British English, where its use is lighter and less likely to be used for situations to describe serious malcontent. The two methods, to varying degrees, confer on these results.

These results are then compared to those of a stepwise logistic regression analysis using the dialect distinction as a response variable. The regression analysis clearly shows that MCA, despite the complexity of the scatter plots, better captures the relative associations revealed in the data. This is most likely due to the need to conflate variables in HCA architecture which in turn may cause low frequency cells. This may explain why the results of the HCA seem to "lump together" less frequent correlations. Although the regression model reveals that the results of the MCA are more informative, the correlation of certain data points is erroneously represented. At one point, a rarely occurring, but crucial, feature is shown to be associated with one dialect, where in fact this merely results from a superficial effect of the two dimensional representation of multidimensional space; the feature in question being "drawn towards" the dialect variable because of its association with another non-relevant data point.

In conclusion, the comparable results of both methods demonstrate their usefulness as exploratory techniques. However, both HCA and MCA can be unreliable when faced with complex multivariate data. In light of this, their results warrant confirmatory analysis. Neverhteless, the contrast in the results of the complicated analysis across the three lexemes, suggest that MCA is better suited to truly multivariate exploratory research. One possible advance for these techniques lies in integrating bootstrap resampling and expectation maximisation. Implementing such algorithms may resolve some of the concerns that these exploratory methods face.

References

Arppe, A. (2006). Frequency Considerations in Morphology, Revisited - Finnish Verbs Differ, *SKY Journal of Linguistics*, 19: 175-189.

Divjak, D. (2006). Delineating and Structuring Near-Synonyms. In *Corpora in cognitive linguistics: corpus-based approaches to syntax and lexis*, St. Gries & A. Stefanowitsch (eds), 19-56. Berlin: Mouton.

Divjak, D. and Gries, St. (2006). Ways of Trying in Russian. *Journal of Corpus Linguistics and Linguistic Theory*, 2: 23-60.

Geeraerts, D. (2006). Methodology in Cognitive Linguistics. In *Cognitive Linguistics. Current Applications and Future Perspectives*, G. Kristiansen & al. (eds), 21-50. Berlin: Mouton.

Geeraerts, D. Grondelaers, S., & Bakema, P. (1994). *Structure of Lexical Variation. Meaning, naming, and context*. Berlin: Mouton.

Glynn, D. (in press). Polysemy, Syntax, and Variation. A usage-based method for Cognitive Semantics. In *New Directions in Cognitive Linguistics*. V. Evans & S. Pourcel (eds). Amsterdam: Benjamins.

Glynn, D. Geeraerts, D., & Speelman, D. (forthcoming). Frames, Fields, and Parasynonymy. Developing usage-based methodology for Cognitive Semantics. In *Cognitive Foundations of Linguistic Usage Patterns.* H.-J. Schmid & S. Handl (eds). Berlin: Mouton.

Gries, St. and Stefanowitsch, A. (2006). *Corpora in cognitive linguistics: corpus-based approaches to syntax and lexis*. Berlin: Mouton.

Gries, St. (2001). A multifactorial analysis of syntactic variation: particle movement revisited. *Journal of Quantitative Linguistics*, 8: 33-50.

Gries, St. (2003). *Multifactorial analysis in corpus linguistics: a study of Particle Placement*. London: Continuum.

Gries, St. (2006). Corpus-based methods and cognitive semantics: The many senses of to run. In *Corpora in Cognitive Linguistics*, S. Gries & A. Stefanowitsch (eds). 57-100. Berlin: Mouton.

Gries, S. T. & David, C. (forthcoming). This is kind of/sort of interesting: variation in hedging in English. In *Towards multimedia in corpus linguistics*. P. Pahta & al. (eds). Helsinki: University of Helsinki.

Grondelaers, S., Geeraerts, D., Speelman, D. 2007. A Case for Cognitive Corpus Linguistics. In *Methods in Cognitive Linguistics.* M. Gonzalez-Marquez & al. (eds), 149-170. Amsterdam: Benjamins.

Heylen, K. 2005. A Quantitative Corpus Study of German Word Order Variation. In *Linguistic Evidence: Empirical, Theoretical and Computational Perspectives*, S. Kepser & M. Reis (eds), 241-264. Berlin: Mouton.

Heylen, K, Tummer, J. & Geeraerts, D. (in press). Methodological issues in corpus-based Cognitive Linguistics. In *Cognitive Sociolinguistics. Language Variation, Cultural Models, Social Systems*. G. Kristiansen & R. Dirven (eds). Berlin: Mouton.

Tummers, J., Heylen, K., & Geeraerts, D. (2005). Usage-based approaches in Cognitive Linguistics: A technical state of the art. *Corpus Linguistics and Linguistic Theory*, 1: 225-261.

Tummers, J, Speelman, D. & Geeraerts, D. (2005). Inflectional variation in Belgian and Netherlandic Dutch: A usage-based account of the adjectival inflection. In *Perspectives on Variation. Sociolinguistic, Historical, Comparative*. N. Delbecque & al. (eds), 93-110. Berlin: Mouton.

Wulff, S. (2003). A multifactorial corpus analysis of adjective order in English. *International Journal of Corpus Linguistics*, 8: 245-82.

Harald Hammarström
Chalmers University
harald2@cs.chalmers.se

Basic Word Order Frequencies and Transition Probabilities in the Languages of the World

Traditionally typologists look at frequencies of various types of languages of the world to gain insight about possible human languages. At least potentially, this reflection might be skewed by "historical accidents" that happened to surface as large-scale areal relationships. Whether or not this is an actual problem, one solution to it has already been suggested (i.e., a method to estimate the natural incidence of various types of languages that is [meant to be] immune to historical accidents). Originally proposed by Maslova (2000) and taken up by Cysouw (2007), the idea is to change from estimating probabilities of occurrence to estimating probabilities of transition. At the center of this approach lies the assumption that there is a *constant probilibity of change* inherent in every linguistic parameter, henceforth CPCH ("constant probilibity of change hypothesis"). This further allows the interpretation of frequent types as *stable*, i.e., the constant probability distribution favours changes to the type and disfavours changes from it, versus infrequent types as *less stable*, i.e., the constant probability distribution disfavours changes to the type and favours changes from it (Maslova and Nikitina tted).

In addition to CPCH, the Maslova/Cysouw model also allows birth- and death effects, henceforth BDE ("birth-death effects"). That is, languages, in addition to transitioning in features, can also die and/or fork into to more languages. Thus, languages we find today are not only the result of independent feature transitions from earlier versions of the same languages – they are the surviving members of isolate languages or languages which inherited features from an ancestor language. The specific rates of birth- and death are kept open, but we may assume that birth- and death processes are independent of features. For example, a language is no more (or less) likely to die (or fork) if it has SVO rather than some other value.

We do not question BDE, but we will attempt to show that CPCH is not valid.

We have put together three databases on basic word order:

1. **Ethnologue:** This database contains 1097 data points (Gordon 2005). Sources for the data points are not indicated. It is not clear how the data points/languages were selected.

2. **WALS:** This database contains 1203 data points (Dryer 2005). Sources for the data points are indicated. It is not clear how the data points/languages were selected, but it may be guessed that it is some kind of convenience sample.

3. **Hammarström:** This database contains 338 data points (Hammarström 2007a). Sources for the data points are indicated. The languages were sampled *at random*, one for *every* attested language family in the world.

These three databases put together, without overlap, amount to 2086 languages – possibly the biggest database of a syntactic parameter so far assembled in linguistic typology. Using the classification of Hammarström (2007b), these 2086 languages are fall into 338 distinct families.[1] 198 of the families have only one [language with a] data point (henceforth 'isolates'), and 140 of them have more than one. Intuitively, the word order distribution in the Hammarström sample, the isolates, and the majority word order for the non-isolates, should agree. This property is

[1] According to this classification, a family is a set of languages which have been shown, in publication, using orthodox comparative methodology to be genetically related. This classification, in general, is ignorant of subgrouping matters.

Table 1: Incidence of word order types across samples (see text)

	All 2086		Hammarström		Isolates		Majority	
SOV	977	**46.8%**	208	**61.5%**	121	**61.1%**	86	**61.4%**
SVO	659	**31.5%**	49	**14.4%**	28	**14.1%**	25	**17.8%**
NODOM	166	**7.9%**	30	**8.8%**	17	**8.5%**	11	**7.8%**
VSO	181	**8.6%**	21	**6.2%**	12	**6.0%**	9	**6.4%**
VOS	46	**2.2%**	9	**2.6%**	6	**3.0%**	3	**2.1%**
OVS	14	**0.6%**	6	**1.7%**	3	**1.5%**	2	**1.4%**
VSO/VOS	9	**0.4%**	7	**2.0%**	6	**3.0%**	0	**0.0%**
OSV	13	**0.6%**	1	**0.2%**	1	**0.5%**	2	**1.4%**
SVO/VSO	6	**0.2%**	2	**0.5%**	1	**0.5%**	1	**0.7%**
SOV/OVS	4	**0.1%**	2	**0.5%**	2	**1.0%**	0	**0.0%**
SVO/VOS	6	**0.2%**	0	**0.0%**	0	**0.0%**	0	**0.0%**
SOV/OSV	2	**0.0%**	2	**0.5%**	0	**0.0%**	1	**0.7%**
SOV/SVO	2	**0.0%**	1	**0.2%**	1	**0.5%**	0	**0.0%**
SOV/VOS	1	**0.0%**	0	**0.0%**	0	**0.0%**	0	**0.0%**
	2086		338		198		140	

fully satisfied, as shown in Table 1. The discrepancy to the full 2086-language database is readily understood as distorting effects of a certain few large SVO-prominent families.

The validity of the CPCH may then be assessed by looking at intra-family divergence.

A rigorous statistical test cannot be built because 1) the CPCH is not sufficiently precisely formulated; for example, there are question marks for how much divergence from the constant transition probability is acceptable, and it is not obvious how to quantify time-depth/family-heterogeneity/family-size (or any other transition unit), and 2) the data is not uniformly sampled within families. However, one prediction of the CPCH in any variant, is that the estimates of the constant probilities, if they exist, should become better the larger the family/the more data points we have for the family. For example, if the CPCH gives rise to a stationary distribution of 61% SOV, then we could look at (say) SOV-original families and see how many of its syncronic languages are no longer SOV. If the number of data points for the family is 2, then we expect to find 0.5 or 1.0, but as the number of data points for a family grows, we expect to find incidences that gravitate towards the assumed stationary ratio, in this example 0.61%. More precisely, the logic is as follows:

1. We assume that CPCH is true.
2. Given that PCH is true, it should give rise to a *stationary distribution*.
3. This stationary distribution should be the distribution evidenced above (in the isolates and Hammarström sample).
4. Given the stationary distribution, for each family, we can calculate the maximum likelihood hypothesis of the word order of its ancestor language (we may also note that the ML, MAP and majority vote on a family turns out to give essentially the same results for this data set).
5. Given the ancestor word order and the samples of synchronic word orders attested, we can compare families with the same ancestor word order.
6. Under the assumption that CPCH is true, we expect that families with the same ancestor word order should show similar synchronic distributions. In particular, we expect that the

larger the family/the more data points we have, the synchronic distribution should approach the stationary distribution.

7. We find that the data do not show a converging behaviour.

For space reasons, the full data cannot be given but Table 2, shows synchronic distriutions for the biggest SOV- and SVO-original families respectively. For whatever reason, different language families display very different transition patterns, and there is no observable tendency towards oscillation towards a constant as data points increase. Lexically very diverse families as well as lexically very tight-knit families show divergent rates of word order change.

Intuitively, the presence of BDE introduce some perturbance, to the effect that different families should show diverging behaviour even if CPCH is true. However, we can cope with this, given reasonable assumptions on BDE, mathematical state-of-the-art and that CPCH should be falsifiable at all within practical limits of world's attested languages. We will pay special attention to argue that, for the dataset of this size, the steps outlined above all remain robust in the wake of BDE.

It follows that the CPCH hypothesis, at least for the basic word order parameter, must be rejected or reformulated, though a profitable reformulation appears hard to attain. An introduction of subgrouping distinctions will remain infeasible for a long time ahead, as detailed evidence of subgrouping is much less developed (than mere relatedness demonstration) for most of the world's language families. The same can be said for attempts at a better guess (rather than majority vote) at the diachronic original of a family.

References

Cysouw, M. (2007). Investigating transition probabilities in the world atlas of language structures (wals). Paper Presented at The seventh International Conference of the Association for Linguistic Typology (ALT VII), CNRS, Paris, September 25-28, 2007.

Dryer, M. S. (2005). Order of subject, object, and verb. In B. Comrie, M. S. Dryer, D. Gil, and M. Haspelmath (Eds.), *World Atlas of Language Structures*, pp. 330–333. Oxford University Press.

Gordon, Jr., R. G. (Ed.) (2005). *Ethnologue: Languages of the World* (15 ed.). SIL International, Dallas.

Hammarström, H. (2007a). A genetically stratified language sample for basic word order typology. Paper Presented at The seventh International Conference of the Association for Linguistic Typology (ALT VII), CNRS, Paris, September 25-28, 2007.

Hammarström, H. (2007b). The language families of world: A critical synopsis. Manuscript available at `http://www.cs.chalmers.se/~harald2/language_families_full.pdf` accessed 25 Sept 2007.

Maslova, E. (2000). A dynamic approach to the verification of distributional universals. *Linguistic Typology 4*(3), 307–333.

Maslova, E. and T. Nikitina (Submitted). Stochastic universals and dynamics of cross-linguistic distributions: the case of alignment types. MS available online at `http://www.stanford.edu/~emaslova/Publications/ProbabilityPubl.html`, accessed 11 Feb 2008.

Table 2: Transition probabilities for the biggest SOV-original families (top) and the biggest SVO-original families (bottom).

SOV-Family	n	SOV	SVO	NODOM	VSO	VOS	Other
Sino-Tibetan	172	**91.2%**	**8.7%**	**0.0%**	**0.0%**	**0.0%**	**0.0%**
Indo-European	106	**52.8%**	**33.9%**	**7.5%**	**5.6%**	**0.0%**	**0.0%**
Trans New Guinea	94	**100.0%**	**0.0%**	**0.0%**	**0.0%**	**0.0%**	**0.0%**
Afro-Asiatic	85	**44.7%**	**40.0%**	**5.8%**	**9.4%**	**0.0%**	**0.0%**
Pama-Nyungan	57	**47.3%**	**12.2%**	**31.5%**	**0.0%**	**3.5%**	**5.2%**
Quechuan	42	**100.0%**	**0.0%**	**0.0%**	**0.0%**	**0.0%**	**0.0%**
Uto-Aztecan	31	**41.9%**	**22.5%**	**19.3%**	**12.9%**	**0.0%**	**3.2%**
Omotic	23	**100.0%**	**0.0%**	**0.0%**	**0.0%**	**0.0%**	**0.0%**
Turkic	20	**100.0%**	**0.0%**	**0.0%**	**0.0%**	**0.0%**	**0.0%**
Tupi	18	**50.0%**	**33.3%**	**0.0%**	**5.5%**	**0.0%**	**11.1%**
Uralic	16	**50.0%**	**37.5%**	**12.5%**	**0.0%**	**0.0%**	**0.0%**
Mande	16	**100.0%**	**0.0%**	**0.0%**	**0.0%**	**0.0%**	**0.0%**
Sepik	14	**92.8%**	**0.0%**	**0.0%**	**0.0%**	**0.0%**	**7.1%**
Chibchan	14	**92.8%**	**0.0%**	**7.1%**	**0.0%**	**0.0%**	**0.0%**
Tucanoan	12	**66.6%**	**0.0%**	**8.3%**	**8.3%**	**0.0%**	**16.6%**
Panoan	12	**83.3%**	**0.0%**	**16.6%**	**0.0%**	**0.0%**	**0.0%**
Dravidian	12	**100.0%**	**0.0%**	**0.0%**	**0.0%**	**0.0%**	**0.0%**
Siouan	10	**100.0%**	**0.0%**	**0.0%**	**0.0%**	**0.0%**	**0.0%**
Nakh-Dagestanian	10	**100.0%**	**0.0%**	**0.0%**	**0.0%**	**0.0%**	**0.0%**
Mongolian	10	**100.0%**	**0.0%**	**0.0%**	**0.0%**	**0.0%**	**0.0%**
...							

SVO-Family	n	SOV	SVO	NODOM	VSO	VOS	Other
Austronesian	240	**7.9%**	**64.1%**	**6.2%**	**12.0%**	**5.8%**	**3.7%**
Atlantic-Congo	201	**4.4%**	**91.5%**	**3.9%**	**0.0%**	**0.0%**	**0.0%**
Mayan	50	**0.0%**	**44.0%**	**4.0%**	**26.0%**	**16.0%**	**10.0%**
Austroasiatic	32	**6.2%**	**84.3%**	**6.2%**	**0.0%**	**3.1%**	**0.0%**
Tai-Kadai	21	**4.7%**	**95.2%**	**0.0%**	**0.0%**	**0.0%**	**0.0%**
Central Sudanic	21	**0.0%**	**71.4%**	**28.5%**	**0.0%**	**0.0%**	**0.0%**
Torricelli	9	**0.0%**	**100.0%**	**0.0%**	**0.0%**	**0.0%**	**0.0%**
Miao-Yao	9	**11.1%**	**88.8%**	**0.0%**	**0.0%**	**0.0%**	**0.0%**
Totonacan	5	**0.0%**	**60.0%**	**0.0%**	**40.0%**	**0.0%**	**0.0%**
Songhay	5	**40.0%**	**40.0%**	**20.0%**	**0.0%**	**0.0%**	**0.0%**
North Halmahera	5	**40.0%**	**60.0%**	**0.0%**	**0.0%**	**0.0%**	**0.0%**
Guaicuruan	5	**0.0%**	**80.0%**	**20.0%**	**0.0%**	**0.0%**	**0.0%**
Zaparoan	3	**33.3%**	**66.6%**	**0.0%**	**0.0%**	**0.0%**	**0.0%**
Koman	3	**0.0%**	**100.0%**	**0.0%**	**0.0%**	**0.0%**	**0.0%**
Iwaidjan Proper	3	**0.0%**	**100.0%**	**0.0%**	**0.0%**	**0.0%**	**0.0%**
...							

Holman, Eric W.[1], Søren Wichmann[2], Cecil H. Brown[3], Viveka Velupillai[4], André Müller[5], and Dik Bakker[6] (ASJP Consortium)
University of California, Los Angeles[1], Max Planck Institute for Evolutionary Anthropology & Leiden University[2], Northern Illinois University[3], Justus-Liebig-Universität Giessen[4], Leipzig University[5], University of Antwerp/Lancaster[6]
D.Bakker@uva.nl[6]

Advances in automated language classification

The paper presents a method for the automatic reconstruction of language relationships taking the Swadesh (1955) 100-item word list as a point of departure. However, the method differs from the original lexicostatistical approach in two fundamental ways. First, the comparison between word forms is done by a computer program (ASJP; automated similarity judgment program) on the basis of Levenshtein's (1966) algorithm, resulting in a distance matrix between individual languages. And second, graphic branching structures illustrating language relatedness (family trees) are generated from this matrix by the way of standard software and algorithms originally developed for the use of biologists in studying phylogenetic relationships (Huson 1998). To accommodate wordlists originally published in a variety of more or less simplified orthographies, a special alphabet, called ASJPcode, was devised which makes use of the QWERTY keyboard symbols only. It contains just 34 consonant symbols and 7 symbols for vowels. These symbols are used for phonological segments defined by the most common points and manners of articulation. Rarer segments are represented by the symbol they most closely resemble in terms of point and manner of articulation. See Brown et al (to appear 2008) for details.

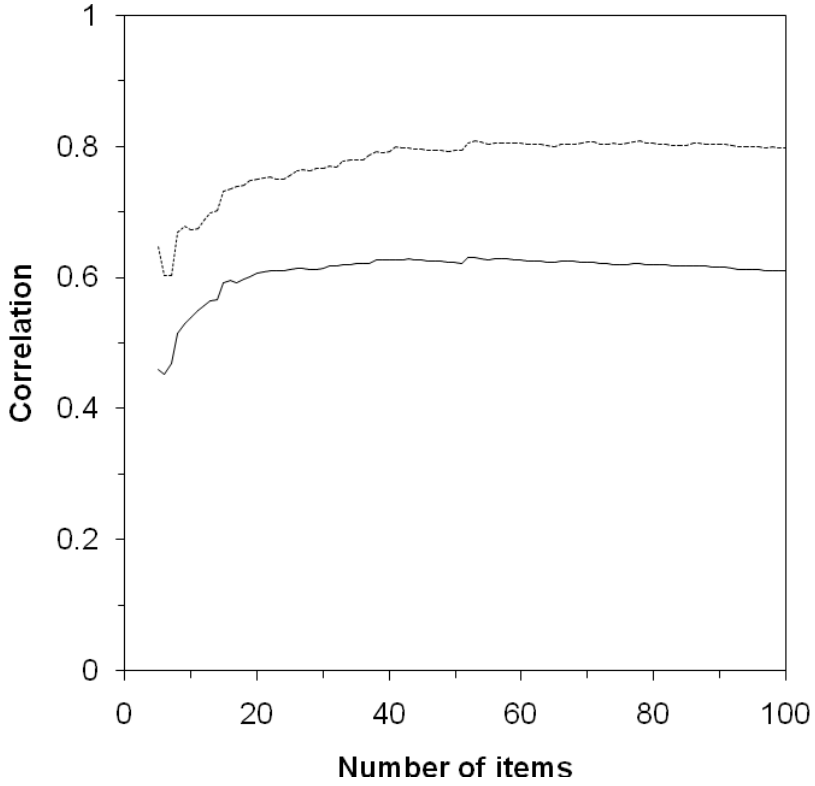
Unlike most other approaches to automatic language classification, such as those described by Oswalt (1971), Atkinson et al. (2005), and Nakleh et al. (2005), the present method automates both the judgments of cognacy and the subsequent inference of phylogeny. We can therefore apply the same objective criteria worldwide to classify an unusually large sample of languages. This facilitates the large scale statistical study of overlaps in lexicons between languages and may reveal previously unknown phylogenetic relationships.

To date, we have collected and transcribed a basic word set for close to 2000 languages of the world. The nearly 2 million language pairs in the database are compared by means of the Levenshtein Distance (LD: see Levenshtein 1966). For any pair of words represented in ASJPcode, LD is defined as the minimum total number of additions, deletions, and substitutions of symbols necessary to transform one word into the other. For any pair of languages L1 and L2, first the LD values are established for each of the N Swadesh words that L1 and L2 share (virtually always the full set that we consider). These LD values are then normalized by dividing each LD by its theoretical maximum giving the normalized LD (LDN). Finally, since lexical similarity may be influenced by chance resemblances, such as an overlap in the phoneme inventories or shared phonotactic preferences for the two languages involved, we correct each LDN by dividing by it the mean LDN of all N(N-1)/2 pairings of words with different meanings, giving the LDND value for each of the N meaning pairs. The LDND value for the

language pair L1 – L2, i.e. their Levenshtein distance, is defined as the mean of the LDND values for the individual word pairs.

Earlier experiments on several hundreds of languages have shown that the 100-item Swadesh list may be reduced to a much shorter one, without loss and even with a gain in classificatory reliability. The subset we selected contains the 40 most stable elements from the original list. Our measure of stability is based on the idea that the more stable items can be identified among a larger set because they have a greater tendency to yield cognates within widely acknowledged groups of closely related languages than words for less stable items. For a comparison of the values in our distance matrix, we have chosen the families and genera as established by Dryer (2005) and the genetic classification of the Ethnologue (Gordon 2005). If we take these classifications as a point of departure, and especially when looking at the more or less firmly and independently established groupings, then the stability factor for the individual lexical items turns out to be consistent across the languages from different hemispheres. Moreover, iterative comparisons lead to a specific subset of 40 items that make better predictions than any smaller subset, and at least as good predictions as any larger subset. The figure below gives an impression of this. A more detailed discussion may be found in Holman et al. (to appear 2008).

The 40-item list contains most of the items in the shorter lists proposed by Yakhontov (see Starostin 1991: 59-60) and Dolgopolsky (1986), and makes better predictions than do the shorter lists.



The ASJPcode was originally introduced for practical reasons: limitations of the keyboard, and problems to represent full IPA code in traditional programming languages.

These two problems have recently been overcome by the project. Full digitalized IPA representations are now automatically converted into equivalent numerical representations that the analyses programs can deal with. Interestingly however, this seems to have no noticeable influence on the results sofar: correlations between the LDN and LDND scores for both IPA and ASJP representations are all significant at the 1% level, and we have noticed no crucial differences between the tree structures produced.

By taking LDND instead of LDN as appoint of departure for further operations, we make an attempt to correct for chance similarities. But no attempt is made to distinguish inheritance from diffusion or universal tendencies. The relative influence of these three factors can be estimated empirically, however, by studying LDND as a joint function of taxonomic distance and geographic distance. For this analysis, geographic distances between languages of the ASJP sample were calculated as the shortest path on the surface of a sphere between the approximate centers of the areas in which the languages are spoken. Comparisons between groups of genetically related and non-related languages show that the amount of similarity declines with distance much more rapidly for the former than the latter groups. This suggests that borrowing of items from the Swadesh list is rather rare between non-related languages, and that most of the weight should be assigned to inheritance. Although there are clear exceptions among the language pairs, our current estimate is that on average not more than 1 or 2 out of the 40 items will be borrowed between non-related languages.

In order to further evaluate the role of lexical comparison we estimated the extent to which acknowledged genetic relationships may be confirmed by other methods. For this purpose we used a subset of the data stemming from the World Atlas of Language Structures (WALS; Haspelmath et al. 2005). Although the WALS project has a purely descriptive goal, and does in no way seek to contribute directly to genetic reconstruction, we think that the wide range and the quality of its database warrants this exercise. So, using the same method as for the Swadesh list to determine the optimal stable subset of the 140 linguistic features of the WALS, we established that for the relatively few languages with at least 100 attested features, the maximum correlations with the Ethnologue and WALS classifications are similar to the correlations for our 40 lexical items in a much larger sample of languages. It follows that equally good results can be achieved either with a high investment of research time in assembling typological features or with a low investment in assembling lexical items.

Furthermore, we studied the behavior of combinations of lexical material and typological features. Our results indicate that fairly close to optimal results are reached using the 40 most stable lexical elements and the 40 most stable typological features for each language, weighted such that lexical elements account for three quarters and typological features account for one quarter of the distance between pairs of languages.

A future goal of the project is to refine the current method of automatically detecting borrowings. It remains to be seen whether for this exercise less abstract representations than the ASJPcode would give better results. An effort will be made, therefore to make full IPA representations available for all languages in the database.

References

Atkinson, Quentin, Geoff Nichols, David Welch, and Russell Gray. 2005. From words to dates: water into wine, mathemagic, or phylogenetic inference? *Transactions of the Philological Society* 103:193-219.

Brown, Cecil et al. (to appear 2008). Automated Classification of the World's languages: A description of the method and prelimary results. *Sprachtypologie und Universalienforschung*.

Dolgopolsky, Aaron B. 1986. A probabilistic hypothesis concerning the oldest relationships among the language families in northern Eurasia. In *Typology, Relationship and Time: A Collection of Papers on Language Change and Relationship by Soviet Linguists*, Shevoroshkin, Vitalij V. and Thomas L. Markey (eds.), , 27-50. Ann Arbor: Karoma.

Dryer, Matthew S. 2005. Genealogical language list. In: Haspelmath et al. (eds.), 584-643.

Gordon, Raymond G., Jr. (ed.). 2005. *Ethnologue*. 15th Edition. SIL International. <www.ethnologue.com>.

Haspelmath, Martin, Matthew Dryer, David Gil, and Bernard Comrie (eds.) 2005. *The World Atlas of Language Structures.* Oxford: Oxford University Press.

Holman, Eric et al. (to appear 2008) Explorations in automated language classification. *Folia Linguistica*.

Huson, Daniel H. 1998. SplitsTree: A program for analyzing and visualizing evolutionary data. *Bioinformatics* 14.10: 68–73.

Levenshtein, Vladimir I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Cybernetics and Control Theory*, 10(8):707-710.

Nakhleh, Luay, Don Ringe, and Tandy Warnow. 2005. Perfect phylogenetic networks: A new methodology for reconstructing the evolutionary history of natural languages. *Language* 81:382-420.

Oswalt, Robert L. 1970. The detection of remote linguistic relationships. *Computer Studies in the Humanities and Verbal Behavior* 3:117-129.

Starostin, Sergei. 1991. *Altajskaja Problema i Proisxozhdenie Japonskogo Jazyka* [The Altaic Problem and the Origin of the Japanese Language]. Moscow: Nauka.

Swadesh, Morris. 1955. Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics* 21: 121-137.

Elsi Kaiser and Jeffrey Runner
University of Southern California, University of Rochester
elsi.kaiser@usc.edu, runner@ling.rochester.edu

Pronouns, reflexives and something in-between: A cross-linguistic investigation of reference resolution in Finnish, German and Dutch

Introduction

In English and in many languages, it has been observed that pronouns and reflexives are in (nearly) complementary distribution. However, the complementarity breaks down in representational NPs (RNPs, e.g. *picture of her/herself*). In English RNPs, (i) interpretation of reflexives is guided by a strong structural subject preference and a weaker semantic *source-of-information* preference (Kuno 1987), and (ii) interpretation of pronouns is guided by a non-subject preference and a *perceiver-of-information* preference (Tenny 2003). These patterns are robust in off-line data and on-line processing (Kaiser et al. 2008), but the nature of the semantic preferences is not well-understood. To further our understanding of the source/perceiver effects, we conducted three experiments investigating the interpretation of pronouns, reflexives and emphatics in RNPs in German, Dutch and Finnish. The experiments aim to shed light on three aspects of the source/perceiver preference: (1) Can the source preference be attributed to intensifiers? (2) Can the source preference be derived from a general prominence bias? (3) How typologically and syntactically robust are the source/perceiver effects? Are they restricted only to certain syntactic constructions or to certain language families?

Question 1: Can the source preference be attributed to intensifiers?

English emphatic intensifiers (ex.1) have the same form as syntactic reflexives (e.g. Koenig & Gast 2006). It has been suggested (e.g. de Vries 1999, see also Bergeton 2004) that intensified object pronouns surface as reflexives (**him himself* => *himself*).

(1) a. Himself used as a reflexive: *The king washed himself.*
b. Himself used as adnominal intensifier: *The king himself opened the doors.*

Thus, reflexives in English RNPs (*picture of himself*) could be proper reflexives or intensified pronouns. If use of intensifiers is guided by semantics (e.g. Koenig & Gast 2006), could the source effects with English RNP reflexives be due to the presence of an intensifier? German can be used to test this: Emphatic intensifiers (*selbst*) are distinct from reflexives (*sich*). If source effects for English reflexives are due to intensification, they should not arise with non-intensifier reflexives. This predicts that in German RNPs, refl+intensifier *sich selbst*, but not plain reflexive *sich*, should prefer sources.

Experiment 1: German

This experiment crossed verb type (*tell/hear*) and anaphoric form (pronoun / reflexive / emphatic), creating six conditions. Participants read sentences (ex.2) and indicated who was shown in the picture (subject/object/either one possible/third person).

(2) *Tobias {erzählte/hörte von} Peter von dem Bild von {ihm / sich / sich selbst}.*
'Tobias {told/heard from} Peter about the picture of {pronoun / refl / emphatic}'

The results of Experiment 1 show that reflexives and emphatics pattern alike: Both preferred the subject (>70%); but this was modulated by a source preference: more subject choices with *tell* (Subj=source) than *hear* (Subj=perceiver), p's<.01. Pronouns trigger more object-choices (overall >50% object-choices, <20% subject-choices, >20% both-choices), but also exhibit a perceiver preference: more object-choices with *tell* (Obj=perceiver, 65%) than *hear* (47%), p's<.01. In sum, the pronoun results resemble English data (see Kaiser et al. 2008). Crucially, since both the plain and the emphatic reflexives prefer sources, the source preference cannot be attributed to an intensifier. This shows that semantic factors must be acknowledged even for plain reflexives.

Question 2: Can the source preference be derived from a general prominence bias?

Existing psycholinguistic research does not explain why pronouns prefer perceivers and reflexives prefer sources. Does this follow from the fundamental distinction between pronouns vs. anaphors/reflexives? We explore another hypothesis, namely that the source preference is due to a general preference for prominent antecedents. Under this view, reflexives' subject preference follows from a preference for structural prominence, and their source preference from a preference for thematic prominence (the sources in Kaiser et al.'s (2008) sentences could be regarded as agentive, see Kuno 1987). If this hypothesis is correct, it predicts that referential forms that prefer prominent antecedents should prefer sources, independently of pronoun/reflexive status. Dutch allows us to test this: Emphatics (pro+intensifier, syntactically pronominal, see de Vries 1999) prefer antecedents that are prominent (de Vries 1999). If the source preference is part of a general prominent antecedent preference, Dutch emphatics should prefer sources. But if pronoun/reflexive status is what determines source/perceiver bias, emphatics (which are pronominal) should prefer perceivers.

Experiment 2: Dutch

The design and methodology were the same as Exp.1. An example sentence is in (3).

(3) *Arne {vertelde/hoorde van} Hans over de foto van {hem / zichzelf / hemzelf}.*
'Arne {told/heard from} Hans about the picture of {pronoun / refl / emphatic}'

Participants' responses reveal that reflexives show an overall subject preference, modulated by a source preference: more subject choices with *tell* than *hear* (78% vs. 63%, p's<.01). Like reflexives, emphatics show a subject preference (50% vs. 18%, p's<.01), and a source preference. However, the subject preference is significantly weaker with emphatics than reflexives (p's<.01). Pronouns trigger approx. 50% both responses (=both subj/obj possible) regardless of verb, but also exhibit a perceiver preference: more object choices with *tell* (35%) than *hear* (17%), p's<.01.

The pronoun-emphatic difference indicates that the source/perceive preference is independent of pronoun/reflexive status, and is compatible with the hypothesis that source preference is part of a general prominence preference.

Question 3: How typologically robust are the source/perceive effects?

Dutch, German and English are all members of the Germanic branch of Indo-European. Do the source/perceiver effects extend to typologically distinct non-Indo-European languages? Furthermore, are these effects restricted to a particular structural configuration? To test this, we investigated whether two kinds of RNPs in Finnish show the same patterns. Finnish has post-nominal RNP constructions (ex.4), similar to English, Dutch and German, but Finnish also has pre-nominal constructions (ex.5) which distinguish pronominal and reflexive-like forms.

In the post-nominal construction, we focus on pronouns, reflexives and emphatics, following the Dutch and German experiments (ex.4). The emphatic form we focus on here is a combination of pronoun+refl (*hänestä itsestään*). Its referential properties are not well-understood; it is not clear whether it is a pronoun modified by an intensifier or a reflexive preceded by an emphatic pronoun (cf. Featherston 2002 for related discussion on German).

In the pre-nominal construction, we tested pronouns, reflexive-like null forms and demonstratives. In Finnish, the presence/absence of the genitive possessive pronoun (*hänen* 's/he-GEN') influences interpretation: it is claimed that an overt possessive pronoun refers to a non-subject (resembling pronouns in English) and its absence (ø in ex.(5a)) indicates subject-reference (resembling reflexives, see Vilkuna 1996). A possessive suffix is present on the head noun in both cases (Vilkuna 1996). To provide a baseline, we also investigated the genitive demonstrative *tämän* 'this-GEN,' which is claimed to prefer non-subjects, similar to *hänen* 's/he-GEN.'

(4) *Mari {kertoi Liisalle / kuuli Liisalta} vitsin {hänestä / itsestään / hänestä itsestään}*
'Mari {told Liisa / heard from Liisa} a joke about {pronoun / ref / emphatic}.'

(5a) *Mari {kertoi Liisalle / kuuli Liisalta} {ø / hänen} muotokuvastaan.*
'Mari {told Liisa / heard from Liisa} about {ø / her} portrait.'

(5b) *Mari {kertoi Liisalle / kuuli Liisalta} tämän muotokuvasta.*
'Mari {told Liisa / heard from Liisa} about this' portrait.'

Thus, Finnish allows us to investigate (i) whether the source/perceiver biases occur in a non-Indo-European language and (ii) whether pre- and post-nominal constructions pattern alike--in particular, whether reflexive elements that are morphologically different (overt reflexives in post-nominal RNPs and null reflexives in pre-nominal RNPs) pattern similarly.

Experiment 3: Finnish

The design and method were basically the same as Exp.1 and 2. Sentences like ex.(4) and (5) were used. The results show that in the **post-nominal construction**, pronouns prefer perceivers-of-information: Participants chose subjects more with *heard* (17%) than *told* (3%), $p<0.01$. However, reflexives and emphatics show no verb effects. With both verbs, reflexives prefer subjects (>90%); emphatics are split between subject and object. In the **pre-nominal construction**, no verb effects arise. Absence of an overt genitive possessor triggers subject-choices (*tell*=99%, *hear*=97%). Demonstratives trigger object-choices (*tell*=94%, *hear*=93%), whereas possessive pronouns show no clear object preference (both *tell* and *hear* result in approx. 60% object choices).

The results for pronouns in Finnish post-nominal RNPs show that the perceiver bias extends to non-Indo-European languages. However, no source preference is observed for regular reflexives or emphatics in post-nominal RNPs. This asymmetry suggests that source effects and perceiver effects can occur independently of each other, a finding which provides further support for the idea that one should not regard these effects as being inherently linked to a constituent's pronominal vs. reflexive status (see Exp.2; we will also discuss briefly the implications of this claim for a more marked compound reflexive form, *omasta itsestään* 'own+refl'). Moreover, the striking absence of any perceiver effects for the pronouns in pre-nominal RNPs suggests that, at least in certain syntactic domains, structural factors can overpower semantic biases that do arise in other syntactic structures: although Finnish pronouns prefer perceivers in post-nominal RNPs, they show no such preference in pre-nominal RNPs.

Conclusions

German and Dutch exhibit a source preference with reflexives and a perceiver preference with pronouns, showing that this phenomenon is not restricted to English. Finnish also shows a perceiver preference with pronouns in post-nominal constructions, extending the results beyond Indo-European and providing further evidence that a purely structurally-oriented approach to anaphor resolution is not sufficient. However, our results make clear that structural factors cannot be disregarded: As the Finnish data indicate, in some syntactic configurations structural factors overrule semantic preferences. We follow Kaiser et al. (2008) in regarding reference resolution as being guided by multiple factors.

In addition, the German data show that a source preference arises with reflexives even when an intensifier is clearly not present. The Dutch data indicate that source/perceiver patterns can be separated from the refl/pro distinction. On a related note, the Finnish results suggest that the source/perceiver opposition should not be equated with the pronoun/reflexive opposition. Put together, these results suggest that, in the languages we investigated, the source preference cannot be blamed on intensification, and instead may be part of a general preference for prominent antecedents. If this approach is on the right track, it provides a potentially promising means of connecting at least some of the seemingly disparate factors that influence anaphor resolution.

References

Bergeton, U. 2004. *The independence of binding and intensification.* PhD diss. USC.

Featherston, S. 2002. Coreferential objects in German. *Linguistische Berichte* 192.

Kaiser, E., J.T. Runner, R.S. Sussman & M. K. Tanenhaus. 2008. The real-time interpretation of pronouns and reflexives. In E. Efner & M. Walkow (eds.), *Proceedings of 37th Annual Meeting of NELS.* GLSA, UMass.

Koenig, E. & Gast, V. 2006. Focused assertion of identity. *Linguistic Typology* 10.

Kuno, S. 1987. *Functional Syntax*. Chicago: University of Chicago Press.

Tenny, C. 2003. *Short distance pronouns, arg. structure and grammar of sentience.* Ms.

de Vries, M. 1999. Het schemergebied tussen pronomina en anaforen. *Nederlandse Taalkunde* 4, 125-160.

Vilkuna, M. 1996. *Suomen Lauseopin Perusteet.* Helsinki: Edita.

Emmanuel Keuleers
Center for Psycholinguistics & CNTS, University of Antwerp
emmanuel.keuleers@ua.ac.be

Predicting exceptions may be harmful

In theories about the representation of inflected forms in the mental lexicon, emphasis is given either to computation or to storage. Theories that emphasize computation see the production of a complex form such as the regular past tense form *walked* as the result of a process that attaches the suffix *-ed* to the stem *walk*, while theories that emphasize storage consider that complex forms are stored in their entirety and that production simply involves retrieval of the full form. However, while theories may disagree on whether a past tense form such as *walked* is retrieved or computed, all theories agree that at least some exceptional forms (e.g., *be–was*, *go–went*) are retrieved rather than computed. To put it more generally, the more exceptional a form is, the more likely its production would be considered the result of a retrieval process.

Nonetheless, being able to *compute* exceptional forms which are assumed to be *retrieved* in ordinary language production seems to be a desirable characteristic for psychologically motivated computational models. In simulation tasks where part of the lexicon is treated as novel material for which complex forms are to be predicted on the basis of the remaining part of the lexicon (henceforth called *lexical reconstruction*), even the correct generation of the most exceptional form counts toward better performance. Similarly, a model that is able to correctly learn the mappings between stems and inflected forms in a particular domain is considered to have fully mastered a skill that is attributed to language users. For instance, Rumelhart and McClelland's (1986) pattern associator for the English past tense was evaluated on its ability to produce inflected forms of existing verbs through a feedforward network.

However, if we would know of a method to cause selective and reversible memory loss in a language user so that we could subject her to a lexical reconstruction experiment, it is highly plausible that her performance would not match that of such a computational model. Specifically, we can assume that she would not generate the attested forms of many exceptional items. Accordingly, it is very doubtful that each time an inflected form is produced, it is generated on the basis of a stem form, as is the case in a pattern associator.

This is not a serious problem if lexical reconstruction actually requires the same abilities as those used by speakers in producing novel complex forms. It may be the case that a model that performs well on lexical reconstruction also performs well on tasks where *pure generalization* ability is tested, i.e., where language users are asked to generate complex forms on the basis of nonce words (also known as *wug* testing, after Berko, 1958). In this paper, I present evidence that models that perform well in a lexical reconstruction task do not perform well in pure generalization task precisely because these models tend to predict *exceptional* patterns while human participants do so to a far lesser degree. This evidence comes from a large-scale simulation study in which Keuleers and Daelemans (2007) specifically contrasted the performance of

computational models on a lexical reconstruction task and on two pure generalization tasks involving Dutch noun plural production. In the lexical reconstruction task models had to predict plural forms for a random selection of 1/20th of the forms in a lexicon of more than 18 000 items. In the first pure generalization task, models had to predict which of two alternative plural forms of a list of nonce nouns was the most frequent choice of participants in an experiment by Baayen et al. (2002); In the second task models had to predict the plural forms for a list of nonce nouns used in an experiment by Keuleers et al. (2007).

The computational models that were used on these tasks were memory-based learning (MBL) models — so called because they make no abstraction from the learning material. These models can be seen as a more sophisticated version of the *k* nearest neighbors approach, in which the class of a novel item is based on the majority class of its *k* most similar neighbors. For instance, in an MBL model with $k = 7$, the plural suffix for a novel noun is based on the most frequent plural suffix among the 7 most similar sounding nouns (technically, all nouns at the 7 nearest distances). Similarity is computed on the basis of aligned phonological representations of these forms. Interestingly, the parameter *k* has a direct relation to the ability of a model to predict exceptional forms. All other things being equal, the lower the value of *k*, the better the model is able to predict inflectional patterns with a low frequency. Accordingly, the higher the value of *k*, the more *exceptional* patterns will be outvoted by more influential patterns.

Figure 1. Prediction accuracy of memory-based learning models with variable *k* on a lexical reconstruction task and on two pure generalization tasks.

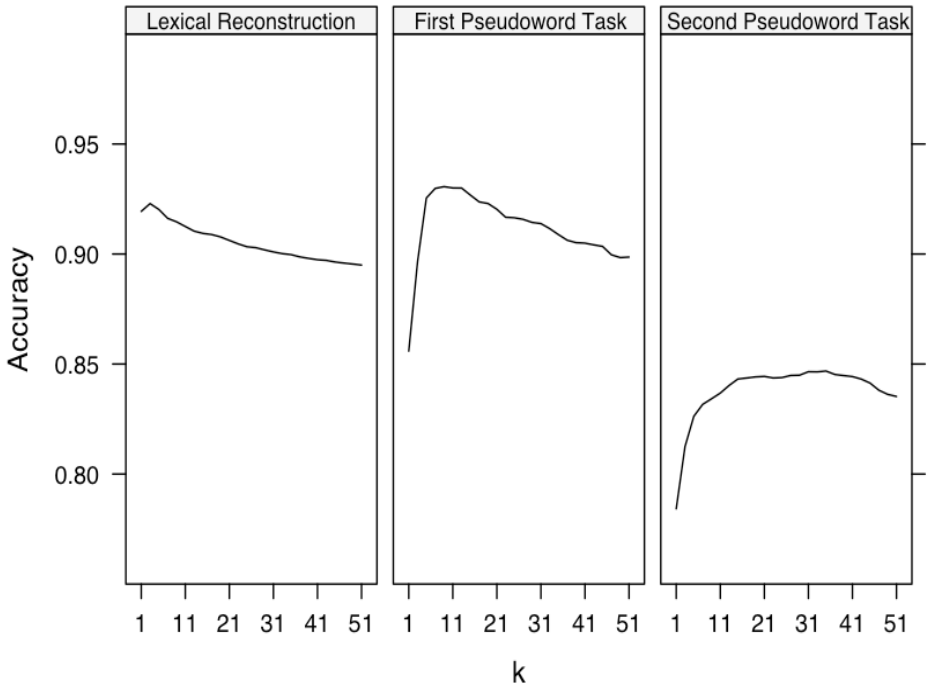


Figure 1 shows the optimal value for *k* for the three tasks described above. While the lexical reconstruction task benefits from a low value for *k*, the pure generalization tasks require a substantially higher value. The most dramatic finding, however, is that a nearly optimal value for the lexical reconstruction task ($k = 1$) is particularly unsuited for the pure generalization tasks. On the basis of these findings, *predicting exceptions* may be considered harmful in modeling language processes. Recent results from memory-based learning of past tense inflection in English (Keuleers & Sandra, submitted) and Dutch (Vandekerckhove, Keuleers, & Sandra, in preparation) support this conclusion. These findings may have consequences on developing theories of language learning. Particularly, they add value to the idea that good performance on the simulation of a linguistic processing task does not necessarily entail similarity to a language user.

References

Baayen, R. Harald, Schreuder, Robert, De Jong, Nivja, and Krott, Andrea 2002. Dutch Inflection: The Rules That Prove the Exception. In *Storage and Computation in the Language Faculty*, Nooteboom, Sieb, Weerman, Fred, and Wijnen, Frank (eds), 61–92. Dordrecht: Kluwer.

Berko, Jean 1958. The child's learning of English morphology. *Word* 14: 150–177.

Keuleers, Emmanuel and Daelemans, Walter 2007. Memory-Based Learning Models of Inflectional Morphology: A Methodological Case Study. *Lingue e Linguaggio* 6 (2): 151–174.

Keuleers, Emmanuel and Sandra, Dominiek. Similarity and Productivity in the English Past Tense. Submitted for publication.

Keuleers, Emmanuel, Sandra, Dominiek, Daelemans, Walter, Gillis, Steven, Durieux, Gert, and Martens, Evelyn 2007. Dutch plural inflection: The exception that proves the analogy. *Cognitive Psychology* 54 (4): 283–318.

Rumelhart, David E. and McClelland, James L. 1986. On Learning the Past Tenses of English Verbs. In *Psychological and Biological Models* [Parallel Distributed Processing. Explorations in the Microstructure of Cognition 2], McClelland, James L., Rumelhart, David E., and The PDP Research Group (eds), 216–271. Cambridge, MA: MIT Press.

Vandekerckhove, Bram, Keuleers, Emmanuel, and Sandra, Dominiek. The productivity of inflectional patterns in the Dutch past tense: The roles of distance and analogical support. Manuscript in preparation.

Victor Kuperman[1], Mirjam Ernestus[1,2], R. Harald Baayen[3]
[1]Radboud University Nijmegen, [2]Max-Planck-Institute for Psycholinguistics, [3]University of Alberta
victor.kuperman@mpi.nl, mirjam.ernestus@mpi.nl, harald.baayen@ualberta.ca

Frequency Distributions of Uniphones, Diphones and Triphones in Spontaneous Speech

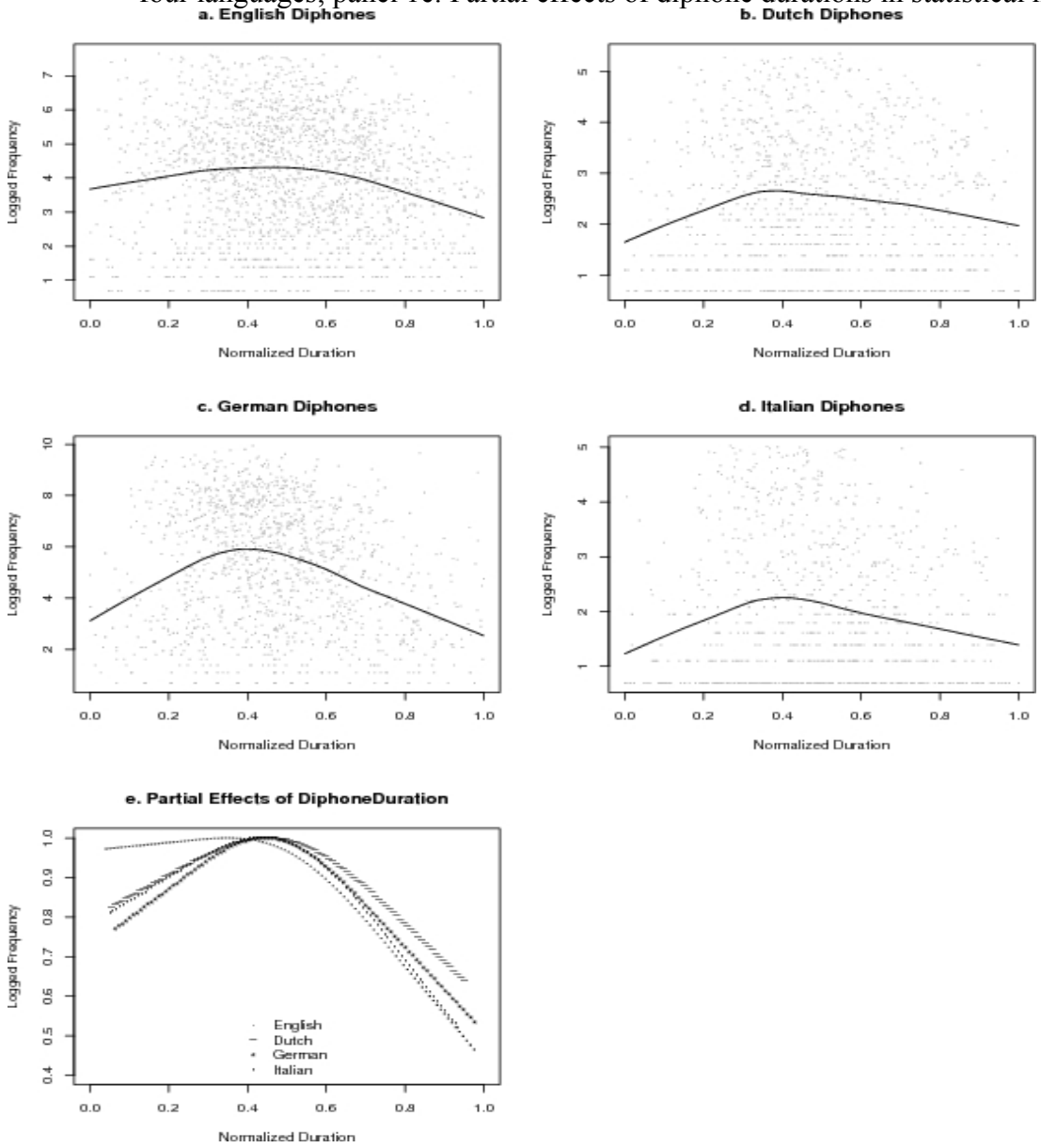
Starting with Zipf (1929; 1935), the overall frequency of occurrence of a speech unit has been argued to enter into a negative (linear or nonlinear) relation with the "degree of complexity" of that unit, which would include its acoustic duration. The more frequent, or otherwise predictable, a speech unit (an n-phone, a syllable, or a word) is, the easier its acoustic realization is claimed to be (cf. Jurafsky et al., 2001). This approach only takes into account the speaker-oriented principle of least effort, but fails to recognize the listener-oriented principle of maximal perceptual contrast as an additional factor that codetermines the relation between frequency of occurrence and production effort. We make the simplifying assumption that acoustic duration of a speech unit reflects (on average, among many other factors) the relative ease of articulating that unit. We hypothesize, along with Zipf (1935), that phonemic sequences with difficult pronunciation will be of a low frequency of use, due to the increased costs for the speaker. In addition, we argue that sequences with extremely easy articulation (e.g., very short ones) may be problematic for the listener and thus be of low frequency in the language as well. The demands of the speaker and the listener may be optimally satisfied by those sequences that are relatively easy to produce and also relatively easy to perceive, that is, by n-phones in the middle of durational range.

In the present paper we tested these hypotheses and explored the relation between frequency of occurrence and acoustic duration of uniphones, diphones and triphones in several languages with different phonemic inventories and different phonologies, namely, English, Dutch, German and Italian. We opted for exploring the relation in spontaneous speech, as several studies show that variation of acoustic duration is larger in this speech variety than, say, in careful speech (e.g,. Johnson, 2004). We based our analyses on large (sub)corpora of spontaneous speech in those languages: The Buckeye Speech Corpus for American English, the IFA Spoken Language Corpus of Dutch, modules Verbmobil-I and -II of the Bavarian Speech Archive for German and the Spoken Italian Varieties Archive for Italian. The speech files of these corpora come with transcriptions at the phone level. Moreover, these transcriptions provide temporal boundaries for each phone in the signal (i.e., phone-level aligned segmentation). Except for the IFA corpus, which was labeled manually, all collections were labeled automatically with subsequent manual verification of the alignment.

We defined diphones (or triphones) as sequences of two (or three) phones without an intervening pause, end of turn, noise, laughter, a non-speech sound, a phone marked as incomprehensible by the transcribers, or a segment extraneous to the phonetic inventory of that language. Notably, in identifying the diphone or triphone sequences we ignored word or utterance boundaries. This approach treats the speech signal as a continuous stream, in which word segmentation is not a given, but rather a task for the listener.

Across the four languages, we found consistent patterns in frequency distributions of diphones and triphones, such that the shortest and the longest n-phones had the lowest frequency of occurrence. In other words, the functional relation between (log-transformed) frequency of occurrence of diphones and triphones as a dependent variable and their (log-transformed) acoustic duration as a predictor, has an inverse-U, concave shape, rather than the monotonically decreasing shape predicted by Zipf's approach, see Fig. 1.

Figure 1, panels a-d: Log frequency of diphones as a function of (normalized) diphone duration across four languages; panel 1e: Partial effects of diphone durations in statistical models.



This set of findings is in line with our hypothesis. For each dataset (e.g., diphones and triphones in each language) we compared the performance of the Zipfian models (that predict a monotonic negative relation) and our models (which predict an inverse-U shape relation). To this end, we used multiple regression models while modeling non-linearities with the restricted cubic splines method. In all cases, our models explained more variance than models based on Zipf's predictions: The average R^2 value of our models was 2.6%, while the average R^2 value of the Zipfian models was 0.2%. The binomial sign test shows that the probability of our models outperforming their counterparts by chance in eight model pairs (four pairs for diphones and four for triphones) is less than 0.008.

N-phone duration can be influenced by a number of factors, including word frequency and speech rate. Can the patterns we observed be explained by those factors? We fitted mixed-effects multiple regression models to each dataset with n-phone duration as a dependent variable, with as fixed effects word frequency, the sum of mean durations of uniphones in the n-phone, mutual information of uniphones, the position of the n-phone in the word and the phrase, and with speaker as a random effect.

We then considered the residuals of those models as a measure on n-phone duration, from which other factors of influence were regressed out. Finally, we considered n-phone frequency as a function of the residual n-phone duration test the performance of our models, and the residual n-phone duration as a function of n-phone frequency to test Zipfian models. The effects of predictors on corresponding dependent variables were statistically significant in all models. Crucially, the advantage that our models showed in fitting the mean durations of diphones and triphones across languages is still preserved when the influence of multiple other predictors is statistically partialled out.

We also tested for whether the inverse-U shape patterns might be an artifact of the so-called sampling error and in fact represent a normal distribution of data points around the mean n-phone duration. For each dataset, we simulated 5000 samples from the normal distribution with the size, mean and the standard deviation equal to those observed in the distribution of residual n-phone durations in the given dataset. The Kolmogorov-Smirnov test invariably showed that the simulated and the observed distributions are significantly different across datasets. We also used the one-sample t-test to estimate the probability that data points in the observed distribution follow the normal distribution (with the mean and standard deviation equal to those of the observed distribution). For over a half (over two-thirds) of data points in each dataset this test showed that their probability of being part of the normal distribution is above the 5% (1%) level of significance. We conclude that the observed distribution patterns cannot be fully accounted for by the statistical fact that values closer to the population mean tend to have higher frequency of occurrence than extreme values.

In order to obtain a better understanding of the observed cross-linguistic patterns, we implemented the hypothesis about the interacting demands of efficient speech production and effective speech comprehension mathematically in a theoretical function based on Job and Altmann (1985). The function is based on assumptions that (a) the relative amount of change in frequency is proportional to the change in the difference in efforts for the interlocutors and (b) language as a self-organization system tends to reach an equilibrium between conflicting processing demands, such as demands of easy production and easy comprehension of speech. The function provides good fits to the distributions of frequency of diphones and triphones over their acoustic durations supporting our hypothesis.

Our data document the existence of consistent frequency distribution patterns in several languages, as revealed via large corpora of spontaneous speech. These patterns demonstrate the emergence of global cross-linguistic regularities from the individual instances of communication that operate on a microscopic scale and provide evidence for processes of self-organization in language.

References

Job, U. and Altmann, G. 1985. Ein Modell fuer anstrenungsbedingte Lautveraenderungen. *Folia Linguistica Historica* , VI:401-407.

Johnson, K. 2004. Massive reduction in conversational American English. In Spontaneous speech: data and analysis. *Proceedings of the 1st session of the 10th international symposium* , pages 29-54, Tokyo, Japan. The National International Institute for Japanese Language.

Jurafsky, D., Bell, A., Gregory, M., and Raymond, W. 2001. Probabilistic relations between words: Evidence from reduction in lexical production. In *Frequency and the emergence of linguistic structure*, Bybee, J. and Hopper, P. (eds.(, pages 229-254. John Benjamins, Amsterdam.

Zipf, G. K. 1929. *Relative frequency as a determinant of phonetic change*. Harvard Studies in Classical Philology , 15:1-95.

Zipf, G. K. 1935. *The Psycho-Biology of Language*. Houghton Mifflin, Boston.

Akira Omaki, Anastasia Conroy, and Jeffrey Lidz
University of Maryland
omaki@umd.edu, staceyc@umd.edu, jlidz@umd.edu

An experimental investigation of referential/nonreferential asymmetries in syntactic reconstruction

Introduction

Syntactic reconstruction effects on reflexive binding (Barss 1986), where the reflexive inside the fronted *wh*-phrase can be bound in either the target (high reading) or base position (low reading) (1), have played a pivotal role in recent discussions of the LF interface (Chomsky 1995, Fox 2000, Sportiche 2006).

(1) $John_1$ wondered which picture of $himself_{1/2}$ $Bill_2$ is likely to hear about *t*.
(cf. $John_1$ wondered if $Bill_2$ is likely to hear about a picture of $himself_{*1/2}$)

One factor that has been argued to condition these binding possibilities is referentiality (Heycock 1995): Reconstruction is optional when the *wh*-argument is referential, but obligatory when it is non-referential. However, the judgment reported in this paradigm is subtle and manipulation of referentiality requires an extremely careful control of the discourse context. We show, using adult data from a variant of a Questions-after-story task (de Villiers, Roeper and Vainikka 1990), that the effect of (non-) referentiality on syntactic reconstruction is much less robust than has been argued in the literature, suggesting that referentiality may not be a crucial factor in determining binding relations.

Previous Observations

Following Heycock (1995), Fox and Nissenbaum (2004) illustrate that referentiality affects the syntactic reconstruction possibilities for reflexive binding. In (2), the semantics of the creation verbs *have ideas* is compatible with a non-referential reading (3a) but not with a referential reading (3b), because the *ideas* are not in existence, and hence cannot be referred to.

(2) How many ideas is John likely to have?
(3) a. What is the number *n* such that John is likely to have *n* ideas?
b. #What is the number *n* such that there are *n* ideas and John is likely to have those ideas?

They claim that when an amount *wh*-phrase contains a reflexive as in (4), then binding in the target position remains available if the *wh*-phrase is selected by a non-creation verb (4a), but not if it is selected by a creation verb (4b).

(4) a. ^{OK}I asked John how many ideas about himself Mary is likely to hear about *t*.
b. *I asked John how many ideas about himself Mary is likely to have *t*.

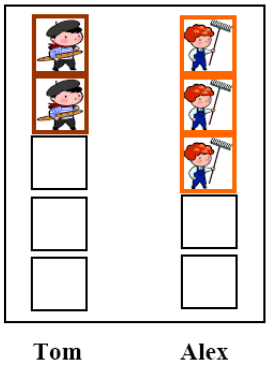
Our experiment tests this contrast, using sentences like (5) with two potential antecedents for the reflexive.

(5) a. Non-creation verb (referential) condition:
Tom wondered how many drawings of himself Alex loved to look at. Do you know?
b. Creation verb (non-referential) condition:
Tom wondered how many drawings of himself Alex needed to draw. Do you know?

Experiment

An experimental test of the contrast presented in (4a) and (4b) is required to control the availability of both referential and non-referential interpretations of the amount *wh*-argument. We created a variant of the Questions-after-story task. In this task, the experimenter presents a scenario that makes available the two potential antecedents for the reflexive. After the scenario, a puppet utters the target sentence followed by a question to the participant (5). The participant answers with a number, from which either the high or low reading of the reflexive in the target sentence can be inferred. For the non-creation verb (referential) condition (5a), the target sentence contains a verb that requires an object in existence (e.g. *look at*), and for the creation verb (non-referential) condition (5b), the target sentence contains a creation verb (e.g., *draw*) that requires a non-existent item. The scenario contains two characters that are posting drawings for an art gallery. There exists a column for each character. The drawings in these columns make available the referential interpretation of the *wh*-argument. The blank boxes, which indicate that more pictures must be drawn to fill those spots, make available the non-referential interpretation of the *wh*-argument (Figure 1).

Figure 1. Example of an art gallery used in our stories



Crucially, each column contains a different number of drawings as well as blank boxes. For example, a response of "2" for (5a) would reflect a high reading of the reflexive in the target sentence. Targets were presented in a pseudorandom order, with control items intermixed. Counterbalanced measures include: number associated with high reading, and side on which the pictures associated with the high reading appeared.

Based on the results from previous self-paced reading experiments and truth-value judgment tasks with adults that used similar sentences, we predicted that

participants would prefer the high reading if both readings are available (Frazier, Plunkett and Clifton 1996; Omaki, Dyer, Malhotra, Sprouse, Lidz and Phillips 2007), but that they would allow only the low reading when the high reading is ungrammatical (Leddon and Lidz 2006). Specifically, if referentiality does condition reconstruction possibilities, we predict that in the non-creation verb (i.e., referential) condition (5a), the subjects would mainly produce high reading answers, whereas in the creation verb (i.e., non-referential) condition (5b), the subjects would not produce high reading answers at all.

Results

Twenty one native speakers of English provided high reading answers 61.9% of the time in the non-creation verb condition (5a), and 69% of the time in the creation verb condition (5b). One-sample t-test shows that the high reading answers were produced at a significantly above chance level (50%) in both non-creation verb condition ($t(20)=6.82$, $p<.0005$) and creation verb condition ($t(20)=7.86$, $p<.0005$), and paired t-test revealed no significant difference between the two conditions ($t(20)= -1$, $p>.1$). These results show that the high reading is not only available, but is preferred in both conditions. This shows that it is not the case that adults require reconstruction when the amount *wh*-argument is selected by a creation verb.

Discussion

The present findings contradict the observation that the syntactic reconstruction of amount *wh*-arguments can be manipulated by the use of creation verbs (Fox and Nissenbaum 2004; Heycock 1995). We present two possible syntactic analyses of these results. Either the referential/non-referential asymmetries in syntactic reconstruction effects are illusory (requiring a carefully controlled, rich discourse context), or the arguments of creation verbs can be rendered referential if they denote objects in *virtual existence* (Sportiche 2006). That is, if the participant interprets (5b) as (7), this *virtual existence* is sufficient in order for an object to be referential, explaining why the high reading was available in our creation verb condition.

(7) Tom wondered for what number x, there are x many drawings of himself Alex needed to draw.
(="Tom wondered how many drawings of himself *there are* that Alex needed to draw.")

This experiment thus reveals intricacies of referentiality that were difficult to capture with traditional syntactic judgments.

References

Barss, A. (1986). *Chains and anaphoric dependencies.* Unpublished doctoral dissertation, MIT, Cambridge, MA.

Chomsky, N. (1995). *The Minimalist Program*. Cambridge, MA: MIT Press.

de Villiers, J., Roeper, T., and Vainikka, A. (1990). The acquisition of long-distance rules. In L. Frazier & J. de Villiers (Eds.), *Language processing and language acquisition* (pp. 257-297). Dordrecht: Kluwer.

Fox, D. (2000). *Economy and semantic interpretation*. Cambridge, MA: MIT Press.

Fox, D., and Nissenbaum, J. (2004). Condition A and scope reconstruction. *Linguistic Inquiry, 35*(3), 475-485.

Frazier, L., Plunkett, B., and Clifton, C. J. (1996). Reconstruction and surface binding. *University of Massachusetts Occasional Papers, 19*, 239-260.

Heycock, C. (1995). Asymmetries in reconstruction. *Linguistic Inquiry, 26*, 547-570.

Leddon, E. M., and Lidz, J. L. (2006). Reconstruction effects in child language. In D. Bamman, T. Magnitskaia and C. Zaller (Eds.), *BUCLD 30 Proceedings* (pp. 328-339). Somerville, MA: Cascadilla Press.

Omaki, A., Dyer, C., Malhotra, S., Sprouse, J., Lidz, J., and Phillips, C. (2007). *The time-course of anaphoric processing and syntactic reconstruction.* Paper presented at the 20th annual CUNY conference on sentence processing.

Sportiche, D. (2006). Reconstruction, binding and scope. In M. Everaert and H. van Riemsijk (Eds.), *The Blackwell companion to syntax* (Vol. IV, pp. 35-93). Oxford: Blackwell Publishing.

Tyler Schnoebelen
Stanford University, Department of Linguistics
tylers@stanford.edu

Measuring compositionality in phrasal verbs

Introduction

This paper demonstrates how to measure the compositionality of phrasal verbs using corpus frequencies from the BNC. This allows us to distinguish semantically transparent phrasal verbs (*they lifted up their hats*) from opaque ones (*they summed up their feelings*). Working by analogy to paradigmatic approaches to morphology (Moscoso del Prado Martín et al 2004), I use information theoretic terms to reveal and express a complicated web of relationships between verbs and particles. In so doing, I am able to predict two different sets of data—one semantic, the other syntactic.

Experiment one: Semantics and parsability

Hay (2002) looked at the ordering of English affixes in terms of their "parsability"—that is, a word like *government* is unlikely to be parsed as *govern+ment* since *government* is more frequent than *govern*. On the other hand, *discern* is more common *discernment*, so the affixed word is likely to be parsed. Similarly, I show that we can determine how distinct the parts of a phrasal verb are by counting the relative frequencies of the verbs participating in phrasal verbs. The prediction, which is borne out, is that literal phrasal verbs will be more obviously made up of parts than opaque ones.

I extract 789 different phrasal verbs from the BNC (3,190 tokens), as well as all tokens of the simplex verbs. Bannard (2002) gives the entailment characteristics of 180 phrasal verbs, and I analyze the parsability of the 124 that are either fully entailed (*lift up* entails both lifting and something going up) or fully unentailed (there is no literal *summing* or *up* in *summing up feelings*). Thinking about phrasal verbs in terms of entailment is a key notion in Bannard (2002); it's also used to good effect in Lohse et al (2004). In a paradigmatic approach like Hay and Baayen (2005), the idea is that forms can get support from other words in the lexicon occupying similar positions. These results suggest that *up* simply isn't as present in *sum up* as it is in *lift up*.

For each phrasal verb in the BNC, I calculate whether or not it was likely to be parsed as a single unit or broken into a verb and a particle by comparing the frequencies of the simplex verb with the verb in its phrasal verb combinations. As long as there are more examples of the simplex verb, the phrasal verb will be parsed. For the verbs that are parsed, I add up how many different "types" there are—this means adding up the number of different particles that they take. For *kick off* we see that its verb combines with not just *off* but *through*, *around*, *up*, and *in*. Thus its "number of types parsed" is 5.

To determine the "average type-parsing ratio" I simply divide the number of parsed types by the total number of types for the verb. There are 18 examples of *wind*; 11 of them with *up*, three of them with *down*, four of them without any particle at all. That means that *wind* has a type-parsing ratio of 1/2=0.5 since *wind down* is parsed but

wind up is not. The bottom two rows in the table are built similarly, only using tokens instead of types.

Table 1. Phrasal verbs behave similar to Hay (2002)'s investigation of affixes.

	Opaque/fully unentailed	Transparent/fully entailed	Significance of difference (by Wilcoxon test)
Avg number of types parsed	2.49	5.52	p=3.76e-06
Avg type-parsing ratio	0.704	0.957	p=0.00336
Avg number of tokens parsed	18.1	33.5	p=0.00156
Avg token-parsing ratio	0.680	0.960	p=0.00336

For each row, it is the transparent column that has the higher value—just as in Hay (2002), where it's the more decomposable/parsable level 2 affixes (*-less*, *-ness*) that score higher than the level 1 affixes (*-ity*, *-ic*). Hay's prediction is that highly parsable affixes "will contain predictable meaning, and will be easily parsed out. Such affixes can pile up at the ends of words, and should display many syntax-like properties" (Hay 2002: 535). Here, in the realm of phrasal verbs, we recall Gries (2002)'s finding that literal items like *lift up* take more advantage of the "actually syntactic" property of flexible alternation between NP objects and particles.

Experiment two: Semantics and information theory

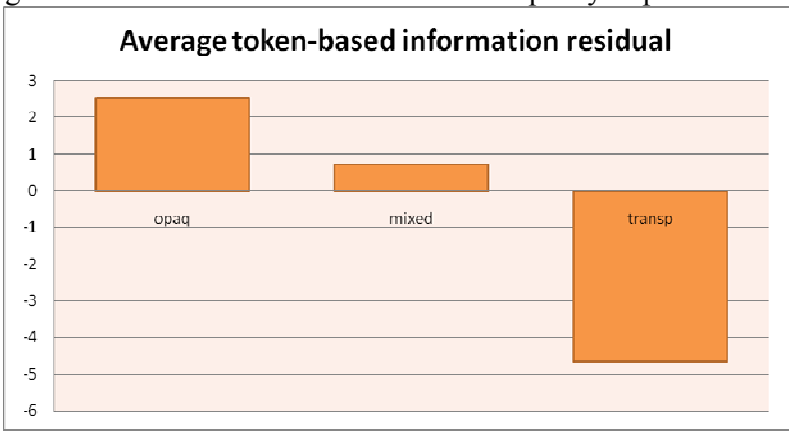
Moscoso del Prado Martín et al (2004) use information theory to develop measures for the amount of information contained in a particular word and the amount carried by the different morphological paradigms it's a part of—in other words, how does a word get composed of meaning? How much does each part and paradigm contribute? Specifically, they calculate the "information residual" based on the overall amount of information ($-\log_2$(frequency of phrase/size of the corpus)) minus the support from its various paradigms, which is measured by a verbal entropy score and a particle entropy score. These numbers are calculated twice—once using token counts and once using type counts. In the type-based calculations, the verb entropy is determined by the number of particles that a verb combines with; the particle entropy is likewise determined by counting how many verbs a particular particle combines with.

Entropy is the number of bits that are necessary to express an outcome—the greater the number of outcomes (and the greater the variety in those outcomes), the greater the entropy (Cover and Joy 2006). Here, there are more outcomes possible for exactly the phrasal verbs that have the largest number of paradigm members: the literal phrasal verbs. Literal phrasal verbs are the most flexible, productive, and intelligible.

There are correspondingly fewer outcomes possible for opaque phrasal verbs, which are more restricted in their meaning and syntax and which are less capable of being parsed into separate pieces. Because the "amount of information" is relatively constant across all phrasal verbs—and because entropy values are subtracted from it—the smaller the entropy values, the larger the information residual. Again, that's the amount of meaning not explained by the parts.

Using 6,793 phrasal verbs, consisting of 2,318 verbs and 48 particles from Baldwin and Villavicencio (2002), I create informational residual scores for all of the phrasal verbs that Bannard (2002) describes. I find that the token-based "information residual" scores for fully unentailed phrasal verbs are reliably higher than that of fully entailed phrasal verbs (p=2.49e-06). The same thing happens in type-based analyses: the informational residual scores for unentailed phrasal verbs are higher than entailed (p=0.01530).

Figure 1. Information residual describes the opacity of phrasal verbs.



Experiment 3: Syntax and information theory

Turning to syntactic realization, I create a generalized linear mixed-effects model with the actual data Gries (2002) used in describing factors that matter for predicting the particle placement of transitive phrasal verbs (*V NP Prt* or *V Prt NP*). Where Gries uses 15 fixed effects, my model has only seven fixed effects and one random effect (the verb itself). Despite the fact that I have simplified the model, I still achieve slightly higher classification accuracy.

Having experimented with no fewer than 26 different variables (including simple log frequency measurements), my final model is comprised of the (i) length of the direct object (DO) in syllables, (ii) the number of times the DO's referent is mentioned in prior discourse, (iii) whether there is a directional adverbial following the DO/particle, (iv) the type of DO (pronominal, lexical, etc.), (v) whether the DO has a definite/indefinite/absent determiner, (vi) the token-based information residual, and (vii) Gries' hand-coded measurement of idiomaticity (idiomatic/metaphorical/literal).

All the factors in the final model are significant and G2 tests demonstrate that removing any of the factors creates a weaker model, while adding any others fails to improve it. This model achieves 87.22% classification accuracy.

Conclusion

While opaque phrasal verbs share the characteristic of "opacity" with idioms, it seems difficult to actually relegate them into the idiom-corner of the lexicon—they don't alternate quite as much or as easily as transparent phrasal verbs, but they still alternate. They also fail other heuristics for idioms (for example, they can passivize). It may be difficult to capture this in grammatical rules unless individual lexical items are marked

and there are different (but very similar) rules that are sensitive to what they find in each lexical entry. Yet even if this approach is tenable, it may not capture the observation that phrasal verbs and their pieces are connected to each other through patterns of usage.

In the first experiment, I used corpus frequencies to demonstrate a difference between semantically opaque and semantically transparent phrasal verbs. The difference lies in the fact that opaque phrasal verbs don't combine with as many particles and the fact that their frequencies, relative to other instances of their simplex verbs, make them more likely to be treated as a single entity.

The next two experiments found the same patterns as the first, but measured them in terms of information theory. While experiment one established that the relationships between particular verbs and particular particles mattered, experiments two and three went further and modeled the relationship between all verbs and particles. By positioning each individual verb and particle in the context of how other verbs and particles were behaving, I showed even stronger results for estimating the entailment characteristics (experiment two) and I was even able to improve models of the "syntactic" phenomena of particle alternation (experiment three).

These corpus experiments establish that analogies to morphology are apt and that it is possible to bring frequencies into syntax and semantics in a meaningful way. Information theoretic terms give us a rich and elegant model for investigating patterns that emerge from actual language use. Such measurements ultimately lead us to ask rather indelicate questions: can generative approaches be adequate if they don't take usage into account? Is compositionality really a categorical phenomenon?

References

Baldwin, T. and A. Villavicencio. 2002. Extracting the unextractable: A case study on verb-particles. In *Proc. of the 6th Conference on Natural Language Learning (CoNLL-2002)*, Taipei, Taiwan.

Bannard, C. 2002. Statistical techniques for automatically inferring the semantics of verb-particle constructions. *LinGO Working Paper No. 2002-06*.

Bolinger, D. 1971. *The phrasal verb in English*. Cambridge: Harvard University Press.

Cover, T. M. and J. A. Thomas. 2006. *Elements of information theory, 1st Edition*. New York: Wiley-Interscience.

Gries, S. T. 2002. *Multifactorial analysis in corpus linguistics: A study of particle placement*. New York: Continuum International Publishing Group Ltd.

Hay, J. 2002. From Speech Perception to Morphology: Affix-ordering revisited. *Language* 78 (3): 527-555.

Hay, J. and Baayen, R. H. 2002. Parsing and Productivity. In *Yearbook of Morphology 2001*, G.E. Booij and J. V. Marle (eds.), 203-235. Kluwer Academic Publishers, Dordrecht.

Lohse, B., J. Hawkins, and T. Wasow. 2004. Processing Domains in English Verb-Particle Constructions. *Language* 80 (2): 238-261.

Moscoso del Prado Martín, F., A. Kostić, and R. H. Baayen. 2004. Putting the bits together: An information-theoretical perspective on morphological processing. *Cognition* 94 (1): 1-18.

Nunberg, G., I. Sag, and T. Wasow. 1994. Idioms. *Language* 70 (3): 491-538.

Dirk Speelman and Dirk Geeraerts
University of Leuven – RU Quantitative lexicology and variational linguistics
Dirk.Speelman@arts.kuleuven.be, Dirk.Geeraerts@arts.kuleuven.be

Putting the (in)direct causation hypothesis to the test: a quantitative study of Dutch *doen* 'make' and *laten* 'let'

In this paper we analyze the choice between the Dutch causative verbs *doen* 'make' and *laten* 'let' in patterns of the form NP CAUSE [NP V (...)] in which CAUSE is a form of either *doen* or *laten*, V is an arbitrary infinite and (...) stands for zero or more constituents which complete the embedded clause. Examples taken from our dataset are given in (1) and (2).

(1) *Ze hebben iemand anders met de caravan laten terugkomen.*
'They have let someone else return with the camper.'

(2) *Als je ze doet tennissen tegen hun zin dan gaan ze niet veel vooruitgang boeken.*
'If you make them play tennis against their will then they will not make much progress.'

The dataset was restricted to cases where it is clear from the context than *doen* 'make' and *laten* 'let' express causation. Most notably, cases where *laten* 'let' expresses permission rather than causation were excluded from the dataset. For instance, in example (1) it was clear from the context that the sentence should be interpreted as 'They have arranged for someone else to return with the camper', and not as 'They have given someone else permission to return with the camper'.

1. Theoretical starting-point

Our theoretical starting-point is the *(in)direct causation hypothesis* that was first formulated by Suzanne Kemmer and Arie Verhagen (Verhagen & Kemmer 1992, Kemmer & Verhagen 1994, Verhagen & Kemmer 1997, Verhagen 1998, Verhagen 2000) and that was more recently analyzed in depth in Ninke Stukker's PhD thesis (Stukker 2005). Drawing on Talmy's notion of force dynamics (Talmy 1988, 2000), the (in)direct causation hypothesis crucially involves the role of the causee in the causative event. The (in)direct causation hypothesis states that the choice for either *doen* or *laten* is influenced by the degree of involvement of the causee. In Stukker's words, in the case of direct causation, as expressed by *doen*, "The causer produces the effected event directly; there is no intervening energy source 'downstream'". In the case of indirect causation, as expressed by *laten*, "Besides the causer, the causee is the most immediate source of energy in the effected event; the causee has some degree of 'autonomy' in the causal process" (Stukker 2005: 50). We will argue in the paper that starting from this assumption about the conceptual difference between *doen* and *laten*, the following more specific hypotheses may be formulated about the distribution of both verbs.

1) If *doen* expresses direct causation, we may expect more *doen* with animate matrix subjects: animate subjects have more control over the flow of energy.

2) If *laten* expresses indirect causation, you don't expect *laten* in constructions with an intransitive infinitive V: in the pattern NP CAUSE [NP V] the second NP typically is the ultimate affectee and the causee is not expressed.
3) If *doen* expresses direct causation, coreferentiality between causer and causee or causer and affectee should favour the use of *doen*: you cannot get more direct than when you exert an influence on yourself.
4) If the relevant factors are purely semantic ones, as in the (in)direct causation model, we don't expect any collocational idiomatization of the distribution: lexical fixation effects should not occur if the distribution is determined by conceptual factors only.
5) At a conceptual level direct causation may be regarded to be the prototypical case of causation, so if *doen* expresses direct causation, its meaning is the center of the causative construction as a whole and can we expect those V infinitives which are themselves typically associated with causative constructions (because of their semantics) to favour *doen*.

2. Dataset and variables

The corpus we used for our case study is the Spoken Dutch Corpus (*CGN - Corpus Gesproken Nederlands*). The Spoken Dutch Corpus (see e.g. Oostdijk 2002 and Schuurman et al. 2003), compiled between 1998 and 2003, contains about 9 million tokens of contemporary spoken standard Dutch. It contains 14 different registers. From this corpus we collected 3975 occurrences of the pattern NP CAUSE [NP V (...)] and we encoded them for the following variables.

The variable cause, with possible values doen and laten, expresses the choice of causative verb and serves as the response variable in the statistical analysis which is discussed in the next section. The following predictors are used to test the specific hypotheses we derived from the general (in)direct causation hypothesis: the variable inanim stands for ‘inanimateness of the first NP'. Its possible values are no and yes. The variable cstr stands for 'construction type'. Its possible values are intransitive and transitive, which stand for intransitive V and transitive V respectively. The variable coref stands for 'coreferentiality'. Its possible values are no and yes, which stand for complete absence of coreferentiality versus presence of some type of coreferentiality respectively. The variable sig.lex.col stands for 'significant lexical collocation' (at an alpha-level of 0.05), and it has two possible values: yes and no. The information we want to store in this variable pertains to 'lexical fixation'. We want to establish whether in some (or many) of the items in our dataset there is (some degree of) lexical fixation at play in the link between the infinitive V and the specific causal verb (either *doen* or *laten*). For this we use a method which is essentially a collostructional analysis (Stefanowitsch & Gries, 2003) although we establish significance by means of the log likelihood ratio test which was introduce into linguistics by Dunning (1993). The variable sig.sem.col is designed to capture 'significant semantic (or conceptual) collocations', as opposed to the more conventional 'significant lexical collocations' captured by sig.lex.col. The variable sig.sem.col is designed to reflect whether there is a significant attraction between the infinitive at hand and the 'abstract causative construction as such' (making abstraction of the specific causative verb). The rationale behind the variable is that verbs which are attracted to the infinitive slot of causative constructions, do so because their meaning easily links up with the concept, i.e. the semantics, of causation. This rather less conventional type of collocation analysis will

be discussed at length in the paper.

Apart from the (in)direct causation hypothesis related variables we also added two variables by means of which we want to verify some additional variationist assumptions. The predictor country, with possible values nl (for The Netherlands) and be (for Belgium) simply encodes whether an observation is drawn from the Netherlandic Dutch or the Belgian Dutch part of the Spoken Dutch Corpus. The predictor spont, with possible values yes and no, simply encodes whether an observation is drawn from the spontaneous speech part (yes) or the prepared speech part (no) of the Spoken Dutch Corpus.

3. Logistic regression analysis

Table 1 lists results from the logistic regression analysis. Variable selection was obtained through forward as well as backward selection (the results were identical). The obtained statistical model is not a simple one since there are some interaction terms (which will be discussed in detail in the paper), but still the overall conclusion must be that several of the (in)direct causation hypothesis induced specific hypotheses were not confirmed by the data, most notably hypotheses 1), 3) and 4).

Table 1: predictor estimates and p values for the logistic regression model

predictors (in order of introduction in forward stepwise regression)	estimates (positive is pro 'doen') and p-values for model with main effects and two way interactions	
(intercept)	-3.26	($p < 0.001$)
inanim (yes)	3.57	($p < 0.001$)
country (be)	1.08	($p < 0.001$)
sig.sem.col (yes)	1.28	($p < 0.001$)
sig.lex.col (yes)	2.33	($p < 0.001$)
sig.lex.col:sig.sem.col	-3.41	($p < 0.001$)
cstr (transitive)	-0.36	($p = 0.25$)
cstr:sig.sem.col	-1.50	($p < 0.001$)
spont (yes)	-0.95	($p < 0.001$)
coref (yes)	-1.23	($p = 0.006$)
inanim:spont	1.23	($p = 0.01$)
cstr:spont	0.67	($p = 0.047$)

4. Interpretation of results

We believe that the case study sheds new light on the (in)direct causation hypothesis. Although this study is no more than a first step towards a thorough quantitative test of that hypothesis, it nevertheless is a substantial one. Although the study does not imply that the hypothesis should be abandoned entirely, it does narrow down the number of legitimate interpretations of the hypothesis. We will argue in the paper that we need to

rethink and refine the (in)direct causation hypothesis on the basis of our findings. We will also suggest an alternative interpretation of the results, which approaches the functional differences between *doen* and *laten* from a different angle.

References

Dunning, Ted 1993. Accurate methods for the statistics of surprise and coincidence. Computational Linguistics 19(1): 61-74.

Kemmer, Suzanne & Arie Verhagen 1994. The grammar of causatives and the conceptual structure of events. *Cognitive Linguistics* 5, 115-156.

Oostdijk, Nelleke 2002. The design of the Spoken Dutch Corpus. In: Pam Peters, Peter Collins and Adam Smith (eds.), *New Frontiers of Corpus Research*, 105-112. Amsterdam: Rodopi.

Schuurman, Ineke, Machteld Schouppe, Heleen Hoekstra and Ton Van der Wouden 2003. CGN, an annotated corpus of spoken Dutch. In: Anne Abeillé, Silvia Hansen-Schirra and Hans Uszkoreit (eds.), *Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora*, 101-108. Budapest, Hungary.

Stefanowitsch, A. and Gries, S.T. 2003. Collostructions: investigating the interaction between words and constructions. *International Journal of Corpus Linguistics* 8.2:209-43.

Stukker, Ninke 2005. Causality marking across levels of language structure. PhD dissertation, University of Utrecht.

Talmy, Leonard 1988. Force dynamics in language and cognition. *Cognitive Science* 12: 49-100.

Talmy, Leonard 2000. *Toward a cognitive semantics*. Cambdrige: MIT Press.

Verhagen, Arie & Suzanne Kemmer 1997. Interaction and causation: Causative constructions in modern standard Dutch. *Journal of Pragmatics* 27, 61-82.

Verhagen, Arie 1998. Changes in the use of Dutch *doen* and the nature of semantic knowledge. In Ingrid Tieken-Boon van Ostade, Marijke van der Wal & Arjan van Leuvensteijn (eds.), *DO in English, Dutch and German. History and present-day variation*, 103-119. Amsterdam/Münster: Stichting Neerlandistiek/Nodus Publikationen.

Verhagen, Arie 2000. Interpreting Usage: Construing the history of Dutch causal verbs. In Michael Barlow & Suzanne Kemmer (eds.), *Usage-Based Models of Language,* 261-286. Stanford, CA: CSLI Publications.

Daphne Theijssen, Nelleke Oostdijk, Hans van Halteren, and Lou Boves
Department of Linguistics, Radboud University Nijmegen
d.theijssen@let.ru.nl, n.oostdijk@let.ru.nl, b.vanhalteren@let.ru.nl, l.boves@let.ru.nl

Modelling the English dative construction in varied written and spoken text

Introduction

Over the past decades, linguistic theorists have attempted to design sets of deterministic rules that account for all-and-only the sentences of a language that are deemed 'grammatical'. However, intuitions about what constructions are (im)possible virtually always appear to be at odds with usage data (e.g. Chater and Manning 2006). The problems related to graded grammaticality and coverage have resulted in probabilistic approaches to linguistics, which consider grammaticality as a function that can take values between 0 (categorically ungrammatical) and 1 (no other option available), with most values being between the two extremes.

There are situations where speakers can choose between several options that are equally grammatical, but that may differ in their acceptability in the given context. An example is the dative construction in English, for which speakers and writers can choose between structures with a double object (NP-NP, e.g. *She handed the student the book*.) or prepositional dative structure (NP-PP, e.g. *She handed the book to the student*.). What we need is yet another kind of descriptive model, which can explain such choices on the basis of a (potentially large) number of linguistic, paralinguistic and extralinguistic properties of a sentence or a paragraph in written texts and their discourse equivalents in spoken language.

Until recently, the development of such models has been hampered by the lack of advanced statistical techniques that can deal with phenomena such as syntactic structures and their elements. Fortunately, linguistics can profit from recent advances in what used to be called nonparametric statistics, where powerful models have been and are being developed for handling this type of variables. The models with which Bresnan et al. (2007) explain the selection between the two dative constructions in English represent arguably the most advanced attempt today to show that the choice between the two options can be explained by way of a combination of (para-)linguistic factors.

In the present research we also aim at modelling the dative alternation, building on Bresnan et al.'s (2007) work. Since their source data is not available due to restrictions on the additions and corrections to the Switchboard Corpus they applied (Bresnan, personal communication), we are forced to create a new data set. This enables us to apply the linguistic features and the statistical modelling techniques they used to data that shows more variation in text genre. Also, we attempt to improve the model by adding new features that we believe are relevant for explaining the variation. Since the research is still in progress, this abstract will only describe our methods, while the results will be presented at the workshop.

Varied written and spoken text

The larger part of Bresnan et al.'s (2007) article concerns transcribed spoken data from the Switchboard Corpus. The model explains 94% of the dative alternation in previously unseen data. They extended the data with instances from the Wall Street Journal texts in the Penn Treebank and concluded that the found model for the spoken data generalizes to written data.

The variety in Bresnan et al.'s data, however, is very narrow. The spoken data contains spontaneous conversations on fixed topics solely, and the written data consists only of financial newspaper articles. Therefore we investigate whether, and if so how, an increase in the range of text and discourse types affects the quality of the model. For this purpose, we employ the syntactically annotated ICE-GB Corpus (Greenbaum 1996). The corpus consists of one million words in British English and contains spoken dialogues (private and public) and monologues (unscripted and scripted), and written texts that are non-printed (student writing and letters) and printed (academic, popular, reportage, instructional, persuasive and creative).

With the help of a Perl script, we automatically extracted sentences with an indirect and a direct object (NP-NP) and sentences with a direct object and a prepositional phrase with the preposition *to* (NP-PP). The found instances have subsequently been manually checked to filter irrelevant structures such as (1a), which contains a locative *to*-PP instead of a prepositional dative construction. For the present research, we ignore constructions with prepositions other than *to*, with coordinated verbs or verb phrases, with phrasal verbs, and with passive voice. Also, we remove all instances with verbs that are present in instances with only one of the two dative constructions. Characteristics of the resulting data set can be found in Table 1.

(1) a. *Fold the short edges to the centre.* (ICE-GB W2D-019_144:1)

b. **Fold the centre the short edges.*

Table 1. Characteristics of our data set

	Spoken		*Written*		*Total*
number of	Dialogues	Monologues	Non-printed	Printed	
texts	180	120	50	150	500
words	360,000	240,000	100,000	300,000	1,000,000
NP-NP	433	222	133	214	1002
NP-PP	84	53	31	52	220
NP-NP / texts	2.4	1.9	2.7	1.4	2.0
NP-PP / texts	0.5	0.4	0.6	0.3	0.4

One of the linguistic features applied by Bresnan et al. (2007) is the semantic class of the verb: 'abstract' (e.g. *give it some thought*), 'transfer of possession' (e.g. *send*), 'future transfer of possession' (e.g. *promise*), 'prevention of possession' (e.g. *deny*) and 'communication' (e.g. *tell*). In the example in the introduction, two noun phrases are important: *the book* (what has been given) and *the student* (who it has been given to). Bresnan et al. call these the 'theme' and the 'recipient', respectively. For both theme and recipient, the discourse accessibility is established as are the pronominality, the definiteness, the animacy, the person, the number and the concreteness (the latter only for the theme). Discourse accessibility is defined as 'given' or 'not given' in the preceding context, or 'accessible' to the addressee. Also, they checked which construction (NP-NP or NP-PP) has been used previously in the dialogue, resulting in the feature 'structure parallelism in dialogue'. Lastly, the length difference between the theme and the recipient is added to the model (log scale). The features 'person of theme' and 'animacy of theme' were removed from Bresnan et al.'s research since they were too sparse. We will follow a similar approach in which we include all features unless they appear to be too infrequent in our data to base conclusions on them. All feature values will be manually determined to reduce the risk of erroneous data.

The statistical modelling techniques Bresnan et al. (2007) apply are Linear Regression Modelling and Generalized Linear Mixed Modelling. The latter is a generalization of the former, in which random effects can be included in the predictor. This results in a model that reveals correlations between the feature effects. Bresnan et al. employ this technique in order to establish the correlation between the verb sense and the other features. We will build similar models for our data set and evaluate the results in comparison with those of Bresnan et al.

Extending the model

Although Bresnan et al. (2007) have based their list of potentially relevant features on a large number of existing theories of and approaches to the dative alternation, we believe there are further linguistic characteristics that are potentially relevant.

Gries and Stefanowitsch (2004), for example, have tried to predict the dative alternation on the basis of the verb form solely. They extracted dative constructions from the ICE-GB Corpus and applied the Fisher exact test to the distribution of each verb form found in both constructions. The results seem promising: for the verb forms with a significant bias towards one of the two constructions (19 of 40), 82.2% of the alternation is correctly predicted, compared to 65.0% when simply selecting the most frequent construction. Therefore, we will include their 'collostructional analysis' in our research as well.

Example (2a) is taken from ICE-GB Corpus, and shows an NP-NP construction in an embedded clause. Although the NP-PP variant we constructed in (2b) is equally grammatical, it is less easy to read and therefore seems less acceptable. This effect can be explained by the principle of end weight, which has also been mentioned in Bresnan et al. (2007). We believe the effect of the principle may increase when the dative construction is embedded deeper in the sentence.

(2) a. *I don't know if a million words would be enough to give [you]* RECIPIENT *[that statistical <,> uhm information to start off with]* THEME. (ICE-GB S1B-076_123:1:B)

b. *I don't know if a million words would be enough to give [that statistical <,> uhm information to start off with]* THEME *[to you]* RECIPIENT.

Having seen instances such as (2), it seems useful to investigate a number of characteristics that relate to the syntactic environment in which the construction is found. Thus, we include information on the level (main or embedded) and type of clause (subordinate or relative), the mode (declarative, interrogative or imperative) and word order (unmarked, clefting or extraposition) of the clause in which the construction occurs, and also information on the polarity (positive or negative) of the clause.

Another feature that does not appear in the feature set of Bresnan et al. (2007) is the presence or absence of an adverb between the theme and the recipient, as exemplified in (3). We will include information on the form and the length of such intervening phrases.

(3) *Ukraine lacks oil, but much Soviet oil comes from the Transcaucasian republics, now also aspiring to independence, which could try to bypass Moscow by selling [oil]*$_{THEME}$ ***directly*** *[to Ukrainian nationalists]*$_{RECIPIENT}$. (ICE-GB W2C-008_20:1)

At the workshop, we will present our results and relate them to the findings of Bresnan et al. (2007) and Gries and Stefanowitsch (2004).

References

Bresnan, Joan, Cueni, Anna, Nikitina, Tatiana, and Baayen, Harald 2007. Predicting the Dative Alternation. In *Cognitive Foundations of Interpretation*, Bouma, Gerlof, Kraemer, Ineke, and Zwarts, Joost (eds.), pp. 69-94. Amsterdam: Royal Netherlands Academy of Science.

Chater, Nick and Manning, Christopher D. 2006. Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences* 10 (7), pp. 335-344.

Greenbaum, Sidney (ed.) 1996. *Comparing English Worldwide: The International Corpus of English*. Oxford: Clarendon Press.

Gries, Stefan Th. and Stefanowitsch, Anatol 2004. Extending collostructional analysis: A corpus-based perspective on 'alternations'. *International Journal of Corpus Linguistics* 9 (1), pp. 97-129.

Bram Vandekerckhove, Emmanuel Keuleers, and Dominiek Sandra
University of Antwerp
bram.vandekerckhove@ua.ac.be, emmanuel.keuleers@ua.ac.be, dominiek.sandra@ua.ac.be

The role of phonological distance and relative support in the productivity of the Dutch simple past tense

Introduction

According to *dual-mechanism* accounts of inflectional morphology, regular inflection is governed by rules that operate over abstract symbols and is therefore fully productive, while irregular inflection depends on a database of stored word forms that allows limited productivity on the basis of similarity-based analogies. According to similarity-based *single-mechanism* models, both regular and irregular productivity depend on analogy-based processing of the target forms. In support of the dual-mechanism hypothesis, Prasada and Pinker (1993) (henceforth P&P) found that people's willingness to produce regular past tense forms for nonce verbs or to give these forms high ratings did not decrease with the nonce verbs' increasing *Phonological Distance* (PD) from existing regular verbs, while ratings and production numbers for irregularly inflected forms did decline with increasing PD of the nonce base forms from existing irregular verbs.

However, in their attempt to design stimuli that differed in their PD from existing irregular and regular verbs, P&P also caused another variable to shift, namely the relative frequencies of the morphological patterns among the closest phonological neighbors of the stimuli, or their *Relative Support* (RS): an increase in PD from existing irregular verbs is accompanied by a rise in RS for the regular inflection while increased PD from regular verbs actually results in a more balanced RS for regular and irregular patterns. This is exemplified by the average number of regular and irregular English verbs that rhyme with the nonce verbs of the different PD classes the authors created (P&P, p. 12).

This finding suggests that the results of P&P should be easy to replicate with a single-mechanism *Memory-Based Language Processing* model (Daelemans & Van den Bosch, 2005) of inflectional morphology. This has been shown by Keuleers and Sandra (submitted) (see also Eddington, 2000). The behavior of such a single-mechanism model is solely determined by the RS for the inflectional patterns among the words in the model's exemplar memory that are closest to the target form. Although the distance between the target word and the words in the lexical memory determines the relative influence of the different lexical items, the global distance between a word and its lexical neighbors itself does not influence the model's inflection choices.

This means, however, that the question whether PD from existing verbs has a *different* effect on regular and irregular productivity is still largely unresolved. We investigated both the effects of PD and RS on the productivity of regular and irregular patterns of the Dutch simple past tense, which like the English past tense system includes one large productive suffixation pattern for regular verbs, and a number of smaller 'gangs' of vowel-change irregulars with limited productivity.

The predictions of three different models were compared: (1) a *Partial-Blocking Dual-Mechanism* model (PBDM), which does not allow any involvement of stored regular items in the productivity of the regular tense but which does allow limited analogy-based productivity of irregular patterns that is able to (partially) block the application of the regular rule, (2) a *Fallback Rule Dual-Mechanism* model (FRDM), which allows both regular and irregular generalization on the basis of lexicon-based analogies, but which uses a symbolic rule for the regular past tense when the lexicon fails to provide a solution because PD becomes too large, and (3) a *Memory-Based Single-Mechanism* model (MBSM), in which both regular and irregular productivity are determined by similarity-based analogies between the target verb and the most similar verbs in the lexicon.

Table 1 summarizes the predictions of the three models. PBDM predicts that there should not be any effect of RS on the productivity of the regular past tense that cannot be accounted for by the number of rule-blocking irregular neighbors. FRDM and MBSM predict a positive effect of RS on the productivity of both the regular and the irregular past tense. In both dual-mechanism models, PD must have a negative effect on the productivity of the irregular inflection, since otherwise the symbolic rule would never be able to come into play: if PD of a word to the items in the lexicon had no effect on their influence, the rule mechanism would never be able to overcome lexical blocking, since the influence of the lexicon would always be equally strong. In PBDM, if PD has any effect at all on the productivity of the regular past tense, it should be a positive one, as increasing PD from the irregular items in the lexicon means less lexical blocking of the symbolic regular rule. It is not entirely clear what PD effect to expect for the regular inflection in FRDM. If there is a noticeable effect of PD on the productivity of the regular pattern it might be a positive one, since the rule operates under increasingly less lexical blocking when PD rises, unless of course rule-generated regular forms receive lower support from the rule mechanism than regular forms that are highly supported by the lexicon. MBSM does not in itself predict a negative effect of PD on the productivity of regular or irregular items, but if it took PD into account, this effect should be negative and equally large for regular and for irregular items when RS is held constant. Both dual-mechanism models predict interaction effects between RS and PD, since the influence of the lexicon should decrease when distance increases. MBSM does not predict such an interaction effect.

Method

Both the effects of PD and RS on the productivity of the Dutch simple past tense were investigated by having participants give acceptability ratings to past tense forms of nonce verbs that varied independently along those two dimensions for the regular

Table 1. Predictions of the three models under investigation concerning the effects of *Relative Support* (RS) and *Phonological Distance* (PD) on the productivity of the Dutch simple past tense.

	Main effect of RS		Main effect of PD		PD/RS interaction	
	Reg	Irr	Reg	Irr	Reg	Irr
PBDM	None	Pos	None/Pos	Neg	Yes	Yes
FRDM	Pos	Pos	None/Pos/Neg	Neg	Yes	Yes
MBSM	Pos	Pos	None (Neg)	None (Neg)	No	No

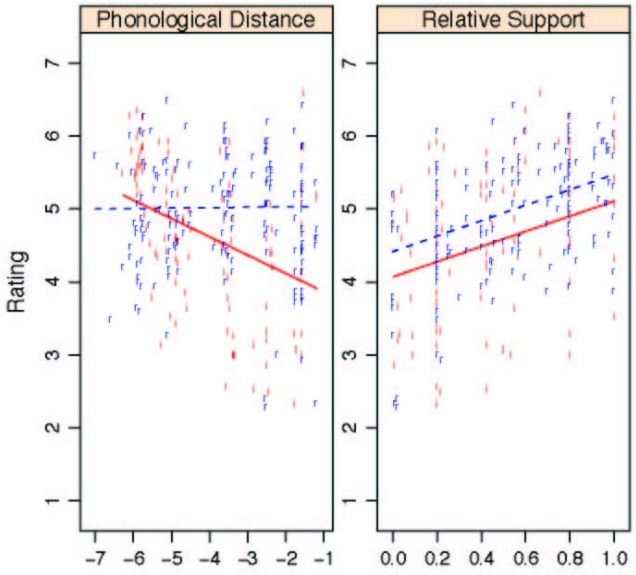
suffixation pattern and the 3 irregular vowel-change patterns with the highest type frequency. 144 monosyllabic nonce verbs were selected as stimuli for this experiment from a large pool of syllables whose phonological representations were assigned to past tense classes using the algorithms from *Tilburg Memory Based Learner* (TiMBL, Daelemans et al., 2008). The training set consisted of the phonological representations of monosyllabic verb stems with their past tense classes, extracted from the Dutch part of CELEX (Baayen et al., 1995). Stimuli were selected by crossing RS for each of the four patterns with mean PD to the nearest neighbors group. Participants were asked to rate the acceptability of simple past tense forms for each nonce verb on a scale from one to seven. 29 undergraduate students from the literature and linguistics department at the University of Antwerp, all native speakers of Dutch, took part in the experiment.

Results and discussion

Mixed effects models of covariance with Participant and Item as crossed random effects were fitted to the ratings for regular forms and irregular forms in stepwise regression analyses. The results (see fig. 1) show equally large positive effects of RS for regular ($\hat{\beta}$ = 1.066, $t(2403) = 6.453$, $p = .0001$) and irregular verbs ($\hat{\beta}$ = 1.04, $t(1735) = 3.658$, $p = .0001$). It does not seem to be the case that this effect of RS on regularly inflected items can be attributed to partial blocking by irregular neighbors, since the correlation between RS and the ratings ($M = 0.54$, $SD = 0.30$, $r(136) = 0.49$, $p = 7.15 \times 10^{-10}$) is significantly higher than that between the ratings and the number of irregular items among the phonological neighbors ($M = 1.29$, $SD = 0.63$, $r(136) = -0.35$, $p = 3.09 \times 10^{-5}$), $Z = -2.33$, $p = .0197$. This seems to rule out PBDM. PD has a significant effect on the ratings for the irregular nonce verbs, $\hat{\beta} = -0.25$, $t(1736) = -4.587$, $p = .0001$. However, there is no significant effect of PD on the ratings for the regular nonce verbs, $\hat{\beta} = 0.0064$, $t(2402) = 0.205$, $p = .857$. MBSM cannot account for this behavior if it does not take PD into account as an independent variable. FRDM does predict this effect of PD. However, this model ideally also predicts an interaction effect between RS and PD, of which we could find no evidence. This means that, although both FRDM and MBSM come close to explaining this pattern of results, the results actually are not straightforwardly explained by any of the models under investigation.

These results lead us to consider some other explanations. The decision which morphological pattern to choose for a given target word might for instance be determined by RS and functions similarly for both regular suffixation and irregular vowel-changes, while in the actual formation of the past tense itself, regular and irregular productivity are differently affected by PD. Another possibility is that participants were very sensitive to the large informative value of the paste tense suffix in rating the past tenses. In an irregularly inflected verb, changing only one of the elements of the forms can have dramatic consequences for its interpretation as a past tense, since the past tense meaning is carried by the whole form. A regularly inflected Dutch verb, however, carries its past tense meaning exclusively on its suffix. This means that one can increase PD to the regular neighbors without diminishing the past tense meaning of the whole verb form.

Figure 1. Partial effects of *Phonological Distance* and *Relative Support* on the ratings for regularly (blue r's, blue broken line) and irregularly (red i's, red full line) inflected nonce verbs.



Conclusions

Although further experiments are necessary to explore all possibilities, these preliminary findings seem to suggest that lexical analogy on the basis of stored regularly inflected verbs plays a crucial role in regular productivity. A partial blocking account in which the application of the regular rule is dependent on the output strength of the memory component does not seem to be able to explain this pattern of results.

References

Baayen, R. Harald, Piepenbrock, Richard, and Gulikers, Leon 1995. The CELEX lexical database (CD-ROM). Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.

Daelemans, Walter and Van den Bosch, Antal 2005. *Memory-based language processing*. Cambridge: Cambridge University Press.

Daelemans, Walter, Zavrel, Jakub, van der Sloot, Ko, and van den Bosch, Antal 2008. *TiMBL: Tilburg Memory Based Learner, version 6.1, Reference Guide.* ILK Research Group Technical Report Series no. 07-07, available from http://ilk.uvt.nl/timbl/.

Eddington, David 2000. Analogy and the dual-route model of morphology. *Lingua* 110(4): 281-298.

Keuleers, Emmanuel and Sandra Dominiek submitted. Similarity and productivity in the English past tense.

Prasada, Sandeep and Pinker, Steven 1993. Generalization of regular and irregular morphological patterns. *Language and Cognitive Processes* 8(1): 1-56.

Amir Zeldes, Anke Lüdeling, and Hagen Hirschmann
Humboldt-Universität zu Berlin
amir.zeldes@rz.hu-berlin.de, anke.luedeling@rz.hu-berlin.de, hagen_h@yahoo.com

What's Hard? Quantitative Evidence for Difficult Constructions in German Learner Data

1. Introduction

Our study is concerned with the identification of 'difficult' structures in the acquisition of a foreign language, which will shed light on theoretical considerations of L2 processing. We argue that – compared to simple vocabulary items or abstract syntactic patterns – structures that contain lexical material as well as categorial variables are especially difficult to acquire. The difficulty level for particular patterns is shown to depend on surface invariability but not on the syntactic categories within which target patterns are embedded. As an example we study the distribution of certain structures which are underused by L2 German learners.

The question "what is difficult for a language learner?" can be addressed using several kinds of data, including learner corpora (e.g. error analysis and over/underuse data, for an overview see Granger et al. 2002), elicitation data, or psycholinguistic studies. Here we focus on corpus data. Previous corpus studies focusing on learner difficulties have examined token and type frequencies in order to calculate vocabulary richness measures, such as lexical density as an index of learner competence (Halliday 1989, Laufer & Nation 1999, and many others). However, lexical frequencies do not tell us what constructions are difficult for learners beyond individual lexemes, nor why. Many other studies (examples are Borin & Prütz 2004 or Westergren-Axelsson & Hahn 2001) focus on interference errors due to the learners' native language (or other learned languages) by comparing learners with a certain L1 to native speakers. Yet in order to establish explanations for difficulties in L2 acquisition independent of a learner's native tongue, we must examine the distributions in native and learner data of e.g. lexemes, collocations, colligations (cf. Stefanowitsch & Gries 2003) and syntactic structures, across learners' linguistic backgrounds. We take the stance that L1-independent underuse phenomena are due to learners either not acquiring patterns, or else avoiding their use despite familiarity with them, in both cases indicating increased difficulty.

2. Data

The data for this study comes from the Falko corpus (**F**ehler**a**nnotiertes **L**erner**ko**rpus des Deutschen als Fremdsprache), which consists of texts from advanced learners of German and control data from German L1-speakers (Lüdeling et al. 2008), allowing contrastive interlanguage analyses. The corpus is stored in a multi-layer model searchable at various levels of annotation. In order to diminish the possibility that the learners are simply unfamiliar with the items in question, we examine only advanced learners and focus on frequent, prevalent patterns. To filter out interference from the learners' L1 and other foreign languages we examine data from speakers of five different L1s: Danish (da), English (en), French (fr), Polish (pl) and Russian (ru), with diverse language education. Using this data, we examine the normalized frequencies of all word form types and part-of-speech *n*-grams in order to find the most significant cases of underuse. Here we focus on two particularly striking cases found in this way, involving reflexives and adverb chains. Use of the reflexive pronoun *sich* can be difficult for learners (Mode 1996), since they must learn not only which verbs and senses require it, but also correctly position it either after the verb in a main clause (1), after a complementizer (2) or subject (3) in subordinate clauses, or initially in an

infinitive phrase (4). Treating the usage of *sich* as a random variable and using a test of equal proportions our data shows very significant underuse of *sich* in both learners in total vs. natives, and each learner dataset grouped by native language vs. natives.

1. *sie entscheiden **sich** meistens für die Firma*
   they decide [refl] usually for the firm
   they usually decide for the firm
2. *dass **sich** die Frauen überfördert fühlen*
   that [refl] the women over-challenged feel
   that the women feel they can't cope
3. *Als die Stadt **sich** ändert*
   as the city [refl] changes
   As the city changes
4. ***sich** ihren Mann auszusuchen*
   [refl] her husband choose
   to choose her (own) husband

L1	natives	learners	da	en	fr	pl	ru
f(*sich*)	.011697	.005910	.006283	.006291	.006930	.007170	.005435
tokens	74280	88736	15593	21600	7786	18100	11203
p-val.		< 2.2e-16	< 3.314e-9	< 8.518e-12	< 1.849e-4	< 1.595e-7	< 3.465e-9

Learners use *sich* about half as often as natives, independent of their L1, even though *sich* is the 17th most common word form in the corpus overall, so it can be assumed that the learners are familiar with it. The examined L1s are quite diverse with regard to the morphosyntax of reflexives (e.g. enclitic or not, position relative to the verb, variability depending on the finite verb's person), yet four of them have similar reflexives (da. *sig*, fr. *se*, pl. *się*, ru. *-sja*). This reduces the likelihood of interference accounting for the underuse phenomenon. Additionally, since interference by definition depends on the learner's native language, we would expect some statistical differences in the underuse patterns between learners with different L1s if interference were a factor (more or less underuse depending on the amount or type of interference). However, the frequency of *sich* in all five learner datasets does not differ significantly (p-val. of .4478 in a 5-way test of equal proportions). Another possible difficulty could be word order complexity of *sich* in relative/infinitive clauses (1-4 above). Yet the data shows *sich* is similarly underused in all syntactic environments, with learner/native normalized frequency ratios of .54 for main clauses, .55 for subordinate clauses and .62 for infinitive clauses, with no significant difference (p-val. of .354 in Pearson's chi-squared test). We therefore conclude that *sich* is similarly underused by our learners independently of their L1 and the embedding clause type.

By contrast, learners do use *sich* more often in certain less variable contexts, such as when the subject is the generic pronoun *man* 'one' (5), despite the fact that *man* itself is not in overuse (an insignificant underuse ratio of ~.95). In these cases the word order in (2) is ungrammatical and only (3) is possible, i.e. *man sich*. This recurring surface pattern is not underrepresented in the learner data (an insignificant underuse ratio of ~.9, cf. row 1 of the table below). Similarly, combinations of *sich* with *lassen* 'allow, let' (6) are also frequent despite an underuse ratio of ~.56 for *lassen*, actually being overused in datasets from three learner L1s and overall (overuse ratio above 1.5 in row 2, though not statistically significant):

5. *Wenn **man sich** bemüht*
   if one [refl] exert
   If one makes the effort

6. *Anhand dieses Beispiels* ***läßt sich*** *erschließen*
   using this example allows [refl] conclude
   Using this example it is possible to conclude

pattern \ L1	learners/ natives	natives	learners	da	en	fr	pl	ru
man* + *sich	.9079	.000563	.000512	.000834	.000509	.001027	.000276	.000089
lassen* + *sich	1.5359	.000078	.000121	.000064	.000185		.000110	.000178
ADV + ADV	.452	.01285	.00581	.01051	.00611	.00616	.00309	.00285
ADV x 3	.265	.00182	.00048	.00109	.00051	.00038	.00011	.00026

In addition to the lexical underuse data above, we also compare frequencies of part-of-speech chains (PoS bigrams and trigrams) in the same corpus. The PoS chains most underrepresented in all examined learner datasets contain two or three consecutive adverbs (and some particles tagged as adverbs, due to the STTS tagset used), with p-value < 2.2e-16 for the bigrams and 1.776e-14 for the trigrams. To explain this phenomenon we examine the 30 most frequent pairs of adverbs qualitatively, since the total amount of chains is too small to evaluate statistically. In order to abstract beyond specific lexical adverb bigrams we divide the chains into four main categories: I. the adverbs belong to different phrases (a 'quasi-pair'; (7) and (8)); or else the adverbs belong to the same phrase which is either II. left-headed (9), III. right-headed (10) or IV. lexicalized (11).

7. *Es ist* [***doch***] [***auch***] *statistisch belegt, dass*
   it is indeed also statistically proven that
   Furthermore, it is indeed statistically proven that
8. *die* (...) *haben* [***schon***] [[***ziemlich*** *viele*] *Lebenserfahrungen*]
   they have already quite many life-experiences
   they already have quite a lot of life experience
9. *ein Kampf, dass bis* [***heute noch***] *andauert*
   a fight that until today still endures
   a fight which has lasted until today
10. *wo es* (...) [[***viel mehr***] *Arbeitsplätze*] *gibt*
    where it much more jobs gives
    where there are many more jobs
11. *und* [***immer noch***] *kann man eine unzufriedenheit spüren*
    and always still can one a discontentment sense
    and still one can sense some discontentment

In category I, we notice a difference between the use of pair types whose elements are sentence- or VP-modifying adverbs (forming two adverbial phrases), as in (7), and those whose second element is a modifier to an adjective phrase or a DP (as an adverbial particle), as in (8). Structures like (7) are very rare in the learner data, whereas structures like (8) seem not remarkably underrepresented. We explain these findings by the different variability of the structures themselves: in (7) the second of the two adverbs (*auch*) can be moved to the initial position of the sentence (*Auch ist es doch statistisch belegt, dass*), or additional elements/phrases can be inserted directly before or after it. Its position is therefore relatively flexible. In (8) *ziemlich* is bound to the adjective phrase with *viele* as a head, it cannot be moved in the sentence without its DP, and no element can be inserted between *ziemlich* and *viele*; its position is fixed. We argue that the differences in frequency are due to differences in the variability of the structures – learners seem to either not acquire topologically flexible elements or be insecure as to where to place them and opt to avoid them.

The single phrase categories II-IV show different patterns. The left-headed phrases

frequent, with too little data to draw any conclusions from. Category III (right-headed phrases like *viel mehr* 'much more' in ex. (10)) is similarly attested for learners and natives, which can be explained by its easy to learn and topologically fixed surface structure. This structure is similar to that of adverbs followed by attributive adjectives (e.g. use of the intensifier *sehr* 'very', in [*eine* [***sehr liebenswerte***] *Gattin*] 'a **very loveable** wife'), which show no statistically significant under- or overuse at all. This may be because their structure is even easier to learn than the one in (8): they have a fixed pattern DP[DET NP[AP[ADV A] N]] with an invariable topological structure.

The lexicalized pairs in category IV (e.g. *immer noch* 'still', ex. (11)) have to be analysed as single units (with no internal structure). Most of these phrases can be at least partly expressed by just one word (*immer noch* → *noch* 'still'), which learners may choose to use instead. We do not find systematic underuse or overuse in all of these cases – such lexical units can apparently be learned like any other, and frequent ones appear to be better represented in many learner groups (e.g. *immer noch* in the Danish, English, and French subcorpora).

3. Discussion

The learner difficulties examined in our study, as identified by underuse statistics, suggest that complex constructions with variable surface forms, such as mobile reflexive pronouns and non-lexicalized adverb chains, hinder effective acquisition of native-like language production. Invariable, frequently recurring patterns, such as lexicalized chains and combinations like reflexive + *man* or *lassen*, facilitate the use of the corresponding constructions. These results conflict with an algebraic model of grammar that might predict that all reflexive verbs and adverb chains are equally likely to be learned, regardless of lexemes (certain adverbs or verbs) or embedded/embedding constructions (*man* as a subject); but they also conflict with models based solely on input frequency. Diverging from the target language distribution, learners seem to filter out reflexives and multiple adverbs in the native usage they are exposed to, but less so when these are embedded in recurrent patterns. This points to a quantitative destructive effect of surface form variability on the learnability of complex structures, possibly connected to processing considerations in the absorption of items in the mental lexicon.

4. References

Borin, L./Prütz, K. (2004) New wine in old skins? A corpus investigation of L1 syntactic transfer in learner language. In: Aston, G./Bernardini, S./Stewart, D. (eds.) *Corpora and Language Learners*, 67-88. Amsterdam: John Benjamins.

Granger, S./Hung, J./Petch-Tyson, S. (eds.) (2002) *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam: John Benjamins.

Stefanowitsch, A./Gries, S. Th. (2003) Collostructions: Investigating the interaction between words and constructions. *Intl. Journal of Corpus Linguistics* 8(2): 209-243.

Halliday, M.A.K. (1989) *Spoken and Written Language*. Oxford: OUP

Laufer, B./Nation, P. (1999) A vocabulary-size test of controlled productive ability. *Language Testing* 16(1): 33-51.

Lüdeling, A./Dolittle, S./Hirschmann, H./Schmidt, K./Walter, M. (2008) Das Lernerkorpus Falko. *Deutsch als Fremdsprache* 2.

Mode, D. (1996) Zur Stellung des Reflexivpronomens *sich* im deutschen Satz. *Deutsche Sprache* 1/96: 34-53.

Westergren-Axelsson, M./Hahn, A. (2001) The use of the progressive in Swedish and German advanced learner English - a corpus-based study. *ICAME Journal* 25: 5-30.

Posters

Renata Enghels
Ghent University
Renata.Enghels@UGent.be

How word order frequencies reveal cognitive schemes: a Romance case study

1. Introduction: word order in infinitive complements

This poster presents the results from an ongoing corpus analysis of word order in Romance infinitive complements (InfC). In the languages considered, namely Spanish, French and Portuguese, this complement type appears with two main verb classes, namely perception verbs (PVs) and causative verbs (CVs). The InfC is a subordinated complement type containing a nominal constituent (NP_2) responsible for the process represented by the infinitive, as illustrated by the following sentences:

Spanish:

(1a) *[…] esperando ver [entrar a un doctor joven y atrevido que le diría, sencillamente: "Vamos"]*$_{\text{InfC}}$. (SOL: Palomino A., 1971)
'[...] waiting to see a young and impudent doctor who would simply say to him: "let's go".'

(1b) *Pero era sólo la lluvia que hacía [crujir las ramas secas del acebuche]*$_{\text{InfC}}$. (CREA: Maqua J., 1992)
'But it was just the rain that made the dead branches of the oleaster crack.'

French:

(2a) *[…] c'est la coutume de voir [le pouvoir échapper à ses détenteurs légaux]*$_{\text{InfC}}$. (LM: 12/2/1994)
'[…] it is costumary to see the power escape from its legal rulers.'

(2b) *Laisse [pousser tes cheveux]*$_{\text{InfC}}$ *[…].* (Frantext: Weyergans F., 1981)
'Let your hair grow […].'

Portuguese:

(3a) *[...] ouvi [um médico dizer para outro que me deviam fazer um TAC [...].]*$_{\text{InfC}}$ (CDP: O Público, 1994)
'[...] I heard one doctor say to another one that they had to make me a CAT scan.'

(3b) *E a porta deixa [passar Cajango e a mulher]*$_{\text{InfC}}$*, um ao lado do outro, […].* (CDP: Amaral M., 1992)
'And the door lets Canjango and his wife pass through, side by side […].'

In Germanic languages such as English and Dutch, the position of the subordinated nominal NP_2 is fixed, always appearing before the infinitive, as showed by the above translations. In Romance languages however, it varies: it can occur before or after the infinitive. This observation leads to the question of why these different word orders exist and by which parameters they are determined.

A comparative/contrastive method is used, since it allows investigating whether the observed correlations are language specific or whether they can be linked to cross-linguistically (though not necessarily universally) valid cognitive schemes. To be more

precise, a quantitative analysis of real discourse examples in three different languages will show that:

(a) word order in the Romance InfC is largely determined by the semantics of the main verb;
(b) this semantics has an impact on the relationships between the main participants (that is, the main subject and the subject of the infinitive complement) of the situation represented by the sentence;
(c) word order can reveal different cognitive schemes or 'dynamicity configurations'.

2. Data collection

Most previous analyses of the syntax of InfCs after PVs (eg. Rodríguez Espiñeira 2000) and CVs (eg. Treviño 1994) lack any empirical foundation, and do not accurately distinguish different verb types (Danell 1979). Therefore, in order to achieve the above-stated goals, the present study builds on a large corpus containing 5732 sentences with InfCs. The category of the PVs is divided between visual (*ver/mirar, voir/regarder, ver/olhar*) and auditory PVs (*oír/escuchar, entendre/écouter, ouvir/escutar*); the class of CVs contains *make*-verbs (*hacer, faire, fazer*) which can be referred to as verbs indicating 'positive causation' as well as *let*-verbs (*dejar/laisser/deixar*) or verbs of 'negative causation' (cf. Soares da Silva 1997). The sentences represent different media of language use (fiction, newspaper articles, etc.) and are taken from electronic databases, namely the *Corpus de Referencia del Español Actual* (CREA, over 150 million words) and the *Corpus del Español* (CDE, 100 million words) for Spanish, *Frantext* (210 million words) and *Le Monde* (1994, 1997-1998) for French and the *Corpus do Português* (CDP, 45 million words) for Portuguese. As table 1 shows the study on PVs has been completed, whereas the conclusions formulated for the CVs are based on a pilot study. Consider the distribution throughout the three languages and the different verb classes:

Table 1. corpus distribution

	PV + InfC			CV + InfC		
	'see'	'hear'	*total*	'make'	'let'	*total*
Spanish	1181	693	1874	100	285	385
French	1700	419	2119	100	290	390
Portuguese	388	376	764	100	100	200

These cases were manually annotated with the following variables: [main verb], [position NP_2], [animatedness NP_1], [animatedness NP_2] and [transitivity Inf]. As to the animatedness of NP_1 and NP_2, we distinguished between human, animate, inanimate self-controlled bodies (the wind, a car,…), inanimate non dynamic and abstract entities. The first three categories were considered to be instances of dynamic participants, whereas the remaining two classes represent non dynamic entities. Finally, the infinitive can also imply different degrees of dynamicity: a semantically transitive verb (such as *eat*) represents a transfer of energy and is thus highly dynamic; an unergative intransitive verb (such as *dance*) represents an emission of energy and is also considered to be dynamic, unlike unaccusative verbs (such as *fall*) which represent a reception of energy by their subject and are thus less dynamic.

3. Discussion

A first quantitative comparison between the three languages shows that for both verb classes – perceptive and causative – the highest number of preverbal NP_2s can be found in Portuguese (PV: 83,7%, CV: 24,5%) and French (PV: 73% - CV: 16%), whereas NP_2 is most frequently postverbal in Spanish (PV: 74,4% - CV: 97%). The following figures illustrate these tendencies:

Figure 1. position NP_2 – perception verbs

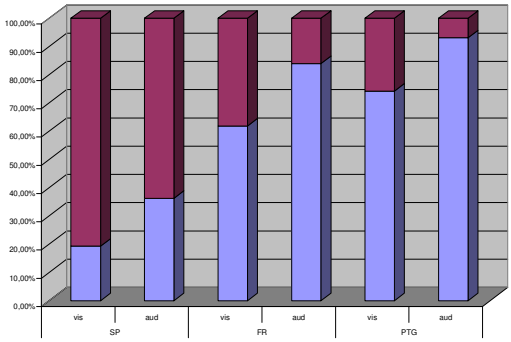
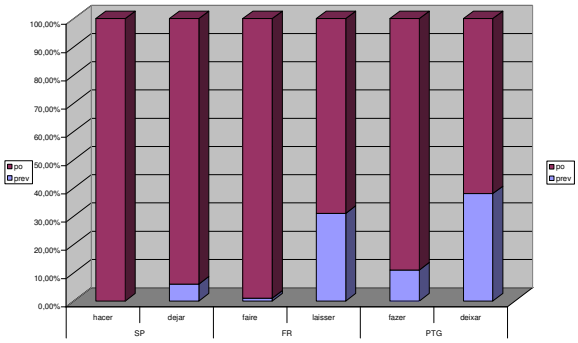


Figure 2. position NP_2 – causative verbs



However, besides these differences the statistical data point towards some striking analogies between the three languages and between the two verb classes. As the above figures show, in Spanish, French and Portuguese:

- auditory PVs more frequently select preverbal NP_2s than visual PVs;
- negative CVs more frequently select preverbal NP_2s than positive CVs.

The main goal of this investigation is to explain these correspondences. It will be argued that what seems to (partly) determine word order in the InfC is the semantics of the main constituents of this complements, which for its part, depends on the extra-linguistic cognitive properties of the perception or causation modality. To be more precise, a thorough corpus analysis will allow us to establish following correlations:

(a) a dynamic NP_2 is more frequently preverbal than a non dynamic NP_2 and NP_2 is more frequently placed before a highly dynamic infinitive, whereas NP_2 mostly appears behind less dynamic unaccusative infinitives;

(b) auditory PVs and negative CVs more frequently select highly dynamic InfCs (with anteposed NP_2s) than visual PVs and positive CVs which more frequently select less dynamic InfCs (with postposed NP_2s).

To conclude it will be demonstrated that these correlations between the main verb type and the dynamicity of InfC depend on the extralinguistic conceptual properties of the perception modalities and the modalities of causative acting.

Firstly, the stimulus of auditory perception needs to produce some noise in order to be heard, whereas the stimulus of visual perception can but does not have to be implicated in an activity. This extralinguistic difference between the two perception modalities explains why in the three languages the auditory PVs mostly opt for dynamic (human, animate non-human and self-controlled) NP_2s and dynamic (transitive and unergative) infinitives, whereas the visual PVs more easily allow less dynamic (inanimate non dynamic) NP_2s and less dynamic (unaccusative) infinitives. Secondly, in positive causation events (‘to make’) the subordinated caused event is mostly dependent

on the main event and non-dynamic, since its occurrence depends on the causative act of the main participant. On the contrary, the subordinated event of negative causation processes is more autonomous and dynamic, since the causer NP_1 opposes to a process that tends to occur any way.

These different dynamicity configurations will allow us to explain the similarities between the three languages on the one hand and the differences within the verb classes on the other one. To put it another way, word order tendencies in the InfC in three Romance languages will be shown to reveal more cross-linguistically valid cognitive schemes.

4. References

Corpus

[CDE] Davies, Mark, Brigham Young University: *Corpus del Español*, http://www.corpusdelespanol.org/

[CDP] Davies, Mark, Brigham Young University / Michael J. Ferreira, Georgetown University: *Corpus do Português*, http://www.corpusdoportugues.org/

[CREA] Real Academia Española: *Corpus de Referencia del Español Actual*, http://www.rae.es/

[FRANTEXT] *Base Textuelle*, http://www.frantext.fr

[LM] *Le Monde*, Cd-rom, 1994, 1997-1998.

Cited references

Danell, Karl Johan 1979. *Remarques sur la construction dite causative: faire (laisser, voir, entendre, sentir) + infinitif*. Stockholm: Almqvist and Wiksell international.

Enghels, Renata 2007. *Les modalités de perception visuelle et auditive: différences conceptuelles et répercussions sémantico-syntaxiques en espagnol et en français*. Beihefte zur Zeitschrift für romanische Philologie. Tübingen: Niemeyer.

Rodríguez Espiñeira, María José 2000. Percepción directa e indirecta en español. Diferencias semánticas y formales, *Verba* 27: 33–85.

Silva, A. Soares da. 1997. *A semântica de deixar: uma contribução para a abordagem cognitiva em semântica lexical*. Lisboa: FCG/MCT.

Treviño, Esthela 1994. *Las causativas del español con complemento infinitivo*. México: El Colegio de México.

August Fenk and Gertraud Fenk-Oczlon
Alps-Adriatic University of Klagenfurt
august.fenk@uni-klu.ac.at, gertraud.fenk@uni-klu.ac.at

Word order and frequency

Introduction

In this contribution we compare three different quantitative studies carried out at the University of Klagenfurt. They use three different of types of methods (A, B, C below) but reveal similar results suggesting the general rule "**more frequent before less frequent**". (A) may be qualified as a corpus based study, (B) is some sort of (hypothesis-guided) text analysis, and (C) an experimental study:

A: Word order in freezes

The respective study (Fenk-Oczlon, 1989) was based on the assumption that the word order in frozen binomials is determined by the rule "more frequent before less frequent" and that this rule would show a higher predictive power than rules such as "short before long", "the first word has fewer initial consonants", "me-first principle", etc. This assumption was tested on 400 freezes from English, Russian and German using the corresponding statistical data (from Thorndike and Lorge, Josselson, Meier, Ruoff). The results include the following:

- With 84% correct predictions the new rule achieves by far the highest accuracy.
- In paired comparisons (all possible combinations of five rules) no other rule achieves such a high degree of correspondence.
- In order to explain those freezes which represent an exception to our rule, recourse must be had primarily to the iconic coding of spatial-temporal relationships.

Jordan (1999) analyzed a total of 579 freezes from French, Italian, and Spanish. In all of these languages our frequency rule showed a higher predictive power than the competing rules.

B: Function words before content words

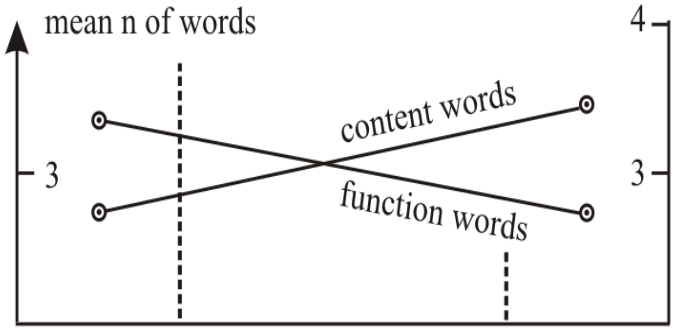
In a previous psycholinguistic study by Auer, Bacik & Fenk subjects were asked to recall as many words as possible from certain sentences of a text by Ernst von Glasersfeld. In a statistical reanalysis (Fenk & Fenk-Oczlon 2006) we found a significantly higher proportion of function words in the primacy part (first quarter) and of content words in the recency part (last quarter) of the sentences. To find out whether this tendency was a characteristic only of the author Glasersfeld we inspected texts – each third of a sentence, if at least 4 words long – from 9 further German authors (4 scientific and 5 literary texts). Table 1 and Figure 1 show the differences between the first and the last quarter of 10 sentences from each of the 10 authors (Fenk-Oczlon & Fenk, 2002).

Table 1: Differences in the distribution of word classes (data material: 10 sentences from each of 10 different German authors)

	1. quarter	4. quarter	diff.
function	3.36	2.67	sign. p<.01
content	2.74	3.46	sign. p<.01
diff.	sign. p<.05	sign. p<.01	

.

Figure 1: The mean number of function words decreases as the mean number of content words increases (data material: 10 sentences from each of 10 different German authors)



Müller (2005) proved these effects in 30 texts (300 sentences) from 3 different Romance languages: In French and Italian the crossing of the two curves to be seen in our Figure 1 shows very late, i.e. near the end of the sentence; in Latin there is no crossing at all. But all of these languages show an increase of content words and a decrease of functions words in the course of a sentence.

C: Behaghels "Gesetz der wachsenden Glieder"

Behaghel (1909) illustrates his law of increasing elements or constituents with many examples from classical texts in a variety of languages such as Ancient Greek, Latin, Old High German and German. In most of his examples the comparison was between word groups of different size or between single words and word groups: *auf der **Türbank** und im **dunklen Gang*** (p.110). In a little experiment by Behaghel the subjects got four sheets of paper with the following words and word groups: *Gold / edles Geschmeide / und / sie besitzt.* They were instructed to form a sentence from these fragments, and the result was always the same: *Sie besitzt **Gold** und **edles Geschmeide***. His explanation: More complex constituents are prepared in the course of sentence production; to place them rather at the end of the sentence meets the cognitive requirements of both, the speaker and the hearer of the sentence. Arnold & Wasow

(2000:28) focus on the role of the hearer when they "argue that postponing heavy and new constituents facilitates processes of planning and production."

In a recent experiment (Fenk & Brunner, 2008) Behaghel's law was tested in a more systematical way, i.e. with varying text material and a higher number of subjects, so that a multiple choice procedure was more appropriate than Behaghel's constructional method. In each of the items the 328 subjects could choose between different sentences such as *Im Labor befanden sich Schafe und Wissenschaftler* versus *Im Labor befanden sich Wissenschaftler und Schafe.* Other than in Behaghel, the test was primarily on the lexical level and was arranged as a competition between Behaghel's "short before long" and our rule "more frequent before less frequent". Each questionnaire contained an equal number of items where the first one of the critical words was short and frequent (a), long and frequent (b), short and rare (c), or long and rare (d).

Assuming that Behaghel's law would also show under these conditions but would be weaker than our frequency rule, the predicted rank order of preferences was a > b > c > d. The respective differences (661 > 615 > 371 >321) as well as the differences between more frequent and less frequent turned out to be highly significant; the difference between short versus long was much lower.

Discussion

The studies (A) and (C) offer a direct comparison between the two regularities "short before long" versus "more frequent before less frequent", and the latter is the clear winner. This shows not only in the binomials (A) where no "hard" syntactical constraints are effective. It seems to be a very robust effect showing in sentences as well: Behaghel found his law, first of all, in texts of different authors from different languages and different periods despite all the syntactical constraints effective in sentence construction. Our experiment (C) exhibits frequency rather than shortness as the relevant factor. Thus both (A) and (C) provide arguments for assuming frequency as the dominant factor, and with respect to (B) we may at least claim that function words tend to be both short and frequent. Therefore the results of all these studies may be subsumed under the covering law "more frequent before less frequent". This rule contributes to a relatively constant flow of linguistic information: The more frequent and thus more familiar elements obtain those initial positions which are *per se* characterized by a higher informational content. That the information, e.g., the uncertainty, is highest in the initial positions of a sequence, is almost trivial from the point of view of information theory and thus also shows in the application of Shannon's (1951) guessing game technique: highest number of errors in the initial positions of sentences, of words, and of syllables (cf. Fenk & Vanoucek 1992: 54). In the course of a sequence the number of errors decreases due to the decreasing number of plausible continuations.

We will argue that the principle of a relatively constant flow of linguistic information is an economy principle and thus incompatible with Croft's (1990: 159) claim that word order is unaffected by such tendencies..

References

Arnold, J.E. and Wasow, T. (2000). Heaviness vs. newness: the effects of structural complexity and discourse status on constituent ordering. *Language,* 76, 28-55.

Auer, L., Bacik, I., and Fenk, A. (2001). Die serielle Positionskurve beim Behalten echter Sätze. Paper presented at the *29. Österr. Linguistiktagung,* 25-27 October, Klagenfurt.

Behaghel, O. (1909). Beziehungen zwischen Umfang und Reihenfolge von Satzgliedern. *Indogermanische Forschungen*, 25, 110-142.

Croft, W. (1990). *Typology and universals*. Cambridge: Cambridge University Press.

Fenk, A. and Vanoucek, J. (1992): Zur Messung prognostischer Leistung. *Zeitschrift für experimentelle und angewandte Psychologie,* 39 (1), 18-55.

Fenk, A. and Fenk-Oczlon, G. (2006). Within-sentence distribution and retention of content words and function words. In P. Grzybek (ed), *Contributions to the Science of Text and Language. Word Length Studies and Related Issues*. Dordrecht: Springer.

Fenk,A. and Brunner, M. (2008). Behaghels Gesetz zur Wortstellung: kurz vor lang oder häufig vor selten? Paper to be presented at the *8. Tagung der Österreichischen Gesellschaft für Psychologie*, 24-26 April, Linz.

Fenk-Oczlon, G. and Fenk, A. (2002). Zipf's tool analogy and word order. *Glottometrics*, 5*,* 22-28.

Fenk-Oczlon, G. (1989). Word frequency and word order in freezes. *Linguistics,* 27, 517-556.

Jordan, M. (1999). *Kommen und Gehen – oder Gehen und Kommen? Die Wortfolge in Binomialen romanischer Sprachen*. PhD dissertation, University of Klagenfurt .

Müller, B. (2005). *Die statistische Verteilung von Wortklassen und Wortlängen in lateinischen, italienischen und französischen Sätzen.* PhD dissertation, University of Klagenfurt.

Shannon, C. E. (1951). Prediction and entropy of printed English. *The Bell Ssytem Technical Journal,* 30, 50-54.

Stefan Grondelaers and Dirk Speelman
Radboud University Nijmegen, University of Leuven
S.Grondelaers@let.ru.nl, dirk.speelman@arts.kuleuven.be

Constructional near-synonymy, individual variation, and grammaticality judgments. Can careful design and participant ignorance overcome the ill reputation of questionnaires?

Background

Few native speakers of Dutch would acknowledge any difference between (1) and (2):

(Dutch)
(1) In de asbak lag *er* een hagelkorrel.
"In the ashtray there was a hailstone"
(2) In de asbak lag een hagelkorrel.
"In the ashtray was a hailstone"

If people would not regard post-verbal *er* "there" in the locative inversion construction as totally superfluous for comprehension, they would have great difficulties glossing its precise contribution to the adjunct-initial sentence. Interestingly, most professional linguists have fared little better with *er* "there", arguably one of the most troublesome words in the Dutch language ever since it was put on the linguistic agenda by Brill in 1854 (*er*'s equivalents in other languages have excited comparable controversy).

What 15 years of data-based investigation has taught us is that *er*'s distribution is multi-factorially and probabilistically motivated. As a result, our task as variation analysts has been to identify *meaningful subgroups* in the data, viz. subgroups of the locative inversion construction which trigger *er* (constructions with temporal adjuncts, with semantically vague locative adjuncts, or with taxonomically unspecific main verbs), but also those sub-varieties of Dutch in which *er* is more frequent (notably Belgian Dutch and informal Dutch). A series of corpus-based regression analyses (Grondelaers, Speelman, & Geeraerts, 2002; 2008; Grondelaers, Geeraerts & Speelman 2007) to which these group factors were added revealed that *er*-preferences in the locative inversion construction can be correctly modelled in about 85 % of all cases. Building on the fact that most *er*-determinants are low-predictability contexts, and inspired by Bolinger's (1977: 92) observation that English *there* signals insufficient contextual anticipation, we conducted a series of self-paced reading and eye-tracking experiments which confirmed that *er* is an *inaccessibility marker*. *Er* is inserted to deactivate inferences which are incompatible with an upcoming low-predictability subject: *ashtray* and *lay* in (1) anticipate "smoked-up tobacco products", not hailstones (Grondelaers, Brysbaert, Speelman & Geeraerts, 2002; Grondelaers, Speelman, Drieghe, Brysbaert & Geeraerts, submitted).

While this function-based predictive success is clearly incompatible with the prevailing idea that *er*'s post-verbal distribution cannot be modelled (De Rooij, 1991), it is interesting to notice that we have never been able to fit *er*'s distribution beyond the 85 % success rate cited above. Observe in this respect that the distribution of the impersonal *il* in French locative inversion constructions can be predicted nearly categorically along similar functional lines (with success estimates going up to 97 %). The inevitable conclusion is that there remains *er*-variation we have not been able to model, either because there are as yet unidentified subgroups of locative inversion constructions or speakers of Dutch which manifest a significantly higher or lower *er*-probability, *or* because there is individual bias. The

latter is not improbable. Individual variation in *er*-preferences was pre-empirically reported in Geerts et al. (1984), De Rooij (1991), Haeseryn et al. (1997), and Van Boxtel (2003). In addition, we have argued that *er* is chosen in Belgian Dutch on the basis of the speaker's *subjective* assessment of the subject's predictability (Grondelaers, Speelman & Geeraerts, 2008), which entails that what is predictable for one speaker or listener need not be predictable for another speaker or listener.

Aim

This paper will, therefore, focus on individual variation, no matter how theoretically and operationally diffuse that concept is. In our regression-based variationist approach to *er*, individual factors are considered as unfittable "noise", as the complement of the group variation which *can* be modelled. In our function-based (psycho-)linguistic approach to *er*, by contrast, individual preferences could be considered as the motivated consequence of the fact that the border between predictable and unpredictable is fuzzy and subjective.

What, then, is the proportion between motivated group variation, motivated individual variation, and non-motivated individual variation (noise)? More specifically, can the proportion of unaccountable "noise" be reduced by a more careful analysis of individual *er*-preferences? A corpus-based answer to this question requires materials in which idiosyncratic uses of *er* (resulting from non-standard assessments of subject predictability) are not eliminated, a condition which excludes virtually all newspaper materials. Since, in addition, sample sizes for spontaneous written or spoken data are too small for reliable analysis, and reliable demographic information is rarely available for these materials, we have no choice but to abandon responsibly collected corpus data, and elicit grammaticality judgments to measure preferential differences between individual listeners.

While introspective judgments continue to represent the standard data collection technique in generative linguistics, they have been under constant attack in other linguistic disciplines for their unreliability and instability (see Schütze 1996; Labov 1996, and Sampson 2007), and for the fact that they are almost never collected according to the standard methodology of psycholinguistic experiments (what raises most concern is the fact that participants are rarely ignorant of the research hypothesis, cf. Wasow & Clark 2005: 1483). The latter authors, however, have convincingly argued that a valid questionnaire design can overcome many of the criticisms against grammaticality judgments. Can we, therefore, develop a rating experiment which makes "predictions about usage which coincide perfectly (…) with what speakers are observed to utter and not to utter in spontaneous speech" (Sampson 2007: 188)? More specifically, can this experiment be designed so carefully that motivated group or individual *er*-preferences are not drowned in unaccountable noise?

Design

A sizeable pool of native speakers (n = 181) rated the grammaticality of 12 short passages containing a locative inversion construction on a 7-point scale. All locative inversion constructions were presented in 2 versions, with and without *er*; participants rated either the version with or the version without *er*. In the 12 critical sentences, three low-predictability factors were orthogonally varied (*temporal* vs. *locative adjunct*, *vague locative* vs *specific locative adjunct*, and *main verb "zijn"* vs. *more specific main verb*). In contrast to previous questionnaire-based approaches to *er*'s distribution (De Rooij 1991 and Van Boxtel 2003), we elicited ratings pertaining to the *global* grammaticality of the passages (not to the appropriateness of *er*), in order to direct attention away from the research question. Critical

passages were presented in two orders to gauge the impact of context on subject predictability and *er*-use.

To check the stability of the ratings, an identical copy of the original questionnaire was administered to the same participants three weeks later. At the end of the second trial, we explicitly asked what participants thought the scientific goal of the experiment had been: 75 % reported ignorance or failed to identify our interest in *er*. Since only 0,94 % of the participants correctly identified *er*'s post-verbal distribution as the exact goal of our enquiry, we can safely state that the absolute majority of participants was ignorant of our research hypothesis.

Results & discussion

Table 1: Linear regression on grammaticality judgments

	Estimate	p-value
Intercept	6,02783	< 2e-16 ***
Adj_vagueloc	-0,41293	1.72e-05 ***
Adj_temp	-0,97596	< 2e-16 ***
Verb_zijn	-1,05573	< 2e-16 ***
Er1	-0,54229	1.17e-05 ***
Prov_antw	-0,12269	0.146156
Prov_limb	-0,32831	0.000299 ***
Prov_ovl	0,08039	0.486234
Prov_wvl	-0,39114	0.000533 ***
Secondtrial	-0.21140	0.000957 ***
Intentionunderstood	-0.08525	0.254584
Adj_vagueloc:zijn	0.61970	2.35e-08 ***
Adj_temp:zijn	-0.32073	0.003831 **
Er:adj_vagueloc	0.39810	0.000330 ***
Er:adj_temp	0.94517	< 2e-16 ***
Er:zijn	0.81111	< 2e-16 ***
Er:prov_antw	-0.02136	0.857994
Er:prov_limb	0.24744	0.053939 .
Er:prov_ovl	-0.56238	0.000578 ***
Er:prov_wvl	0.06754	0.672438
Er:secondtrial	0.17754	0.049817 *
Er:intentionunderstood	0.29690	0.005046 **
multiple R-Squared	0.1533	

A linear regression analysis on the ratings confirms that all interactions between *er* and the low-predictability factors are highly significant: *er* considerably reduces the ungrammaticality experienced when locative inversion constructions contain temporal adjuncts, semantically vague locative adjuncts, or taxonomically unspecific main verbs. While these findings confirm the correctness of the research hypothesis, the interaction "Er:secondtrial", which indicates that *er* is preferred significantly more often ($p = 0.049$) in the second trial of exactly the same questionnaire, strongly suggests that some *er*-variation is not functionally motivated. The low R-Squared (0.1533) raises even more reasons for concern: the evident correctness of the research hypothesis and the careful design of the questionnaire cannot prevent that only a minimal percentage of variation in the grammaticality judgments is motivated by our manipulations. A reliability analysis on the ratings further indicates that a satisfactory

Cronbach's Alpha (> .9) is reached only when all raters (> 40 in each of the 4 conditions) are included in the analysis, which suggests massive individual variation.

The only valid conclusion that can be drawn at this moment is that even when participants are ignorant of the research hypothesis, grammaticality judgments are "too shifty and variable (both from speaker to speaker and from moment to moment)" (Schutze 1996: 3) to reveal much beyond what we already know from other data-collection techniques. Although we have not yet fully analyzed the effect of presentation order – we are currently experimenting with predictability estimates (n-gram probability) to gauge the extent to which the preceding context in the different presentation orders makes subjects more or less predictable –, and although Belgian Dutch is known to manifest more individual variation than Netherlandic Dutch on account of its delayed standardization (Grondelaers et al.: 2008), we fear that the deluge of individual variation observed is technique-related: the inevitable conclusion is that reliable *er*-intuitions cannot properly be elicited in a grammaticality judgment experiment.

References

Bolinger, Dwight. 1977. *Meaning and Form*. London: Longman.

De Rooij, J. 1991. Regionale variatie in het gebruik van er III. *Taal en Tongval* 43, 113-136.

Geerts, G., W. Haeseryn, J. de Rooij & M.C. van den Toorn. 1984. *Algemene Nederlandse Spraakkunst*. Groningen/Leuven: Wolters-Noordhoff.

Grondelaers, Stefan, Marc Brysbaert, Dirk Speelman & Dirk Geeraerts. 2002. *Er* als accessibility marker: on- en offline evidentie voor een procedurele interpretatie van presentatieve zinnen. *Gramma/TTT* 9/1, 1-22.

Grondelaers, S., D. Geeraerts & D. Speelman (2007). A case for a Cognitive corpus Linguistics. In Gonzales-Marques, M., I. Mittelberg, S. Coulson & M. J. Spivey (eds.), *Methods in Cognitive Linguistics*, 149-169. Amsterdam/Philadelphia: John Benjamins.

Grondelaers, Stefan, Dirk Speelman & Dirk Geeraerts. 2002. Regressing on *er*. Statistical analysis of texts and language variation. In Annie Morin & Pascale Sébillot (eds.), *Proceedings of the 6th International Conference on the Statistical Analysis of Textual Data*, 335-346. Rennes: Institut National de Recherche en Informatique et en Automatique.

Grondelaers, Stefan, Dirk Speelman, Denis Drieghe, Marc Brysbaert & Dirk Geeraerts (submitted). Indefinite reference processing: Converging on- and offline evidence for predictive inferencing and remedial cueing. Submitted to *Acta Psychologica.*

Grondelaers, Stefan, Dirk Speelman & Dirk Geeraerts. 2008. National variation in the use of *er* "there". Regional and diachronic constraints on cognitive explanations. To appear in Gitte Kristiansen & René Dirven (eds.), *Cognitive sociolinguistics: Language variation, cultural models, social systems*. Berlin: Mouton de Gruyter.

Labov, W. 1996. When intuitions fail. In L. McNair, K. Singer, L. Dolbrin, and M. Aucon (eds.), *Papers from the Parasession on Theory and Data in Linguistics*, 77-106. Chicago: Chicago Linguistic Society.

Sampson, Geoffrey R. 2007. Grammar without grammaticality. *Corpus Linguistics and Linguistic Theory* 3/1, 1-32.

Schütze, Carson T. 2006. *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology*. Chicago: Chicago University Press.

Van Boxtel, S., P.-A. Coppen & T. Bongaerts (2003). Veel is (er) nog onduidelijk gebleken. Factoren in de keuze voor vervangende subjecten in het Nederlands. *Nederlandse Taalkunde* 8, 181-198.

Milena Jakić[1], Aleksandar Kostić[1] and Dušica Filipović-Đurđević[1,2]
[1]Laboratory of Experimental Psychology, Faculty of Philosophy, University of Belgrade,
[2]Department of Psychology, Faculty of Philosophy, University of Novi Sad
xmile@eunet.yu, akostic@f.bg.ac.yu, dmfilipo@f.bg.ac.yu

The Influence Of The Word Connection Type On The Facilitation Effect In The Lexical Decision Task

Introduction

The results of numerous studies indicate that word recognition is faster when a target word is preceded by the associatively or semantically connected prime word (cf. Meyer & Schvaneveldt, 1971; Koriat, 1981; Neely, Keefe, & Ross, 1989; Schelton & Martin, 1992; Thompson-Schill, Kurtz & Gabrieli, 1998). The aim of this study is to answer two questions: a) is there a facilitation effect that derives from the type of word relation over and above the effect of associative connection, and b) is the facilitation effect between associatively connected words symmetrical, or put differently, will the effect change if we change the positions of the prime and target stimulus. In order to answer these questions we performed two experiments in which we examined the facilitation effect among two groups of word pairs: pairs in which the connection is purely associative and pairs with an associative and semantic type of connection (synonymy, antonymy, hyponymy). In order to compare different types of relations among the stimuli we have chosen the theory of lexical semantics: *componential analysis*. Experiments will also test the predictions of this theory.

Method

Twenty-eight participants (Experiment 1) and twenty-seven participants (Experiment 2) were tested in Lexical decision task (lexical priming paradigm). All the participants were first-year psychology students from the University of Belgrade. The pairs of stimuli were presented on the computer screen (SOA period was 750 ms), and the participants had to decide if the second stimulus was a word of Serbian language or not (choosing a YES/NO button). The dependent variable was the reaction time, which was measured from the beginning of the presentation of the second stimulus until the response.

In each experiment 100 pairs of stimuli were presented, half of which were pseudoword-targets. Both word- and pseudoword-targets were preceded either by neutral context (*****)[1] or by the words (counterbalanced in Latin square design). Word pairs were taken from the list of associative norms, i.e. *The Associative Dictionary of Serbian Language* (Piper, Dragićević & Stefanović, 2005), while the word-primes for pseudoword-targets were chosen to be unassociated to any of word-targets. In the first experiment the word targets were stimuli from the word-association test, while the primes were their most frequent[2] associates. In the second experiment primes and targets

[1] Since the strong facilitation effect is obtained even if the target is preceded by the unrelated word, we decided to measure the facilitation effect of the related word comparing to the neutral context, which should represent the reaction time for a discrete target word.

[2] Frequency of associates was taken from the test of free associations (Piper et al, 2005), in which 800 students took part. Most frequent associates were the words that most participants gave as the response to the stimulus given in a test, and the frequency is the number of participants that had given it.

reversed their positions. There were five groups of associatively connected stimuli pairs. Three of them were also semantically connected: synonymy (*kuća – dom* meaning *house – home*), antonymy (*noć – dan* meaning *night – day*) and hyponymy (*jabuka –voćka* meaning *apple – fruit*). Two groups were not semantically connected: a stronger contextual (*majmun – banana* meaning *monkey – banana*) and weaker contextual connection (*svađa – tašta* meaning *quarrel – mother-in-law*). The difference between the stronger and weaker contextual connection was based upon the associative frequency, whose average values were significantly different. The criteria for the selection of the stimuli concerning the relation of synonymy and antonymy was based upon the primary lexical definitions from the dictionary of Serbian language (Rečnik MS). Since the criteria for the hyponymy was more clear than the criteria for the synonymy and antonymy relation, it wasn't necessary to base the choice of the stimuli on the primary lexical definitions.

Results and discussion

The analysis of variance established the 57 ms effect of primed context comparing to the neutral: $F(1, 49)=47{,}414$, $p<0.0001$. The rest of the analyses were performed on the facilitation effect which was calculated by subtracting reaction times for the target preceded by neutral context and the target preceded by related prime. In table 1 the average values for the five types of relation between primes and targets are given. It is obvious that the five groups of stimuli are very different with regard to distributional statistics, but this couldn't be avoided because of the criteria of stimuli selection.[3] However the differences between the groups were partialed out by means of the analysis of covariance. The analysis of covariance performed on items showed that the facilitation effect was significantly stronger when in addition to associative connection word pairs were also semantically connected (the average difference was 35 ms; figure 1): ($F(1, 43)=7.03$, $p<0.01$).

Table 1. Average values for the stimuli in first experiment

relation type	rating of associative connection	target freq	prime freq	associative prime freq	Rt on target (ms) in neutral context (*****)	Rt on target in related context (ms)	Facilitation effect (ms) ANOVA	Facilitation effect (ms) ANCOVA[4]
Antonymy	4,669	1317	1209	200	584	532	52	61
Hyponymy	5,741	47	57	110	627	562	65	55
Strong context	5,994	397	622	145	605	568	37	40
Weaker context	4,370	367	150	79	638	615	23	25
Synonymy	5,380	122	452	139	732	624	108	91

[3] Besides the fact that there is a limited number of synonyms, hyponyms and antonyms in language, it was also necessary to find the confirmation of strong associative connection between them (in Associative dictionary).

[4] Notice that facilitation effect ANCOVA is not simply obtained by subtracting reaction time on target preceding neutral and connected prime, but also by partialing out (in analysis of covariance) factors known to affect facilitation (frequency of prime and target, associative frequency of prime, rated associative connection, length (number of graphemes) of the target).

The results of the first experiment also indicated that the relation type between words (synonymy, antonymy, hyponymy, stronger contextual and weaker contextual

Figure 1. Average facilitation effect for different type of connections

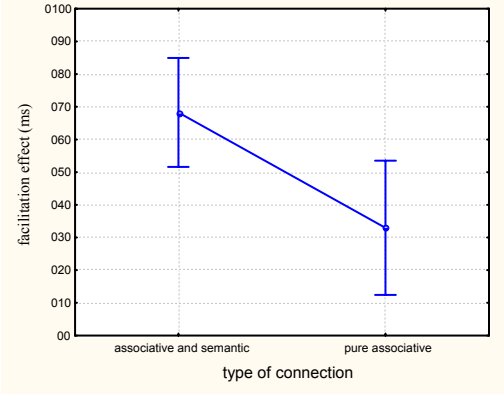
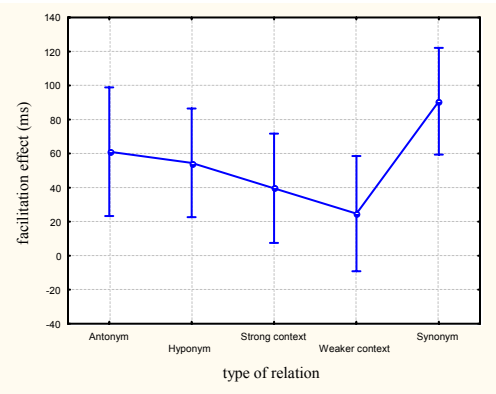


Figure 2. Average facilitation effect for different types of relation



connection) has a significant influence on the facilitation effect: $F(4,40)=2.62$, $p<0.05$ even if we partial out factors known to affect facilitation (lexical frequency of the prime and target (taken from the Kostić, 1999), associative frequency of prime (taken from Piper et al. 2005), target length, rated associative relation between prime and target) figure 2). The strongest facilitation effect was in the group of synonymy. Note that the average lexical frequency for the targets of the synonymy group is higher than average lexical frequency for the targets of the hyponymy group, while the effect is much higher for the synonymy group. On the other hand, the effects of hyponymy and antonymy groups are almost the same, but the average frequency of targets is dramatically different (see Table 1). These outcomes indicate that the observed effects were not due to target frequency.

The differences in the facilitation effect among the five experimental situations, observed ordinally, can be accounted for in terms of *componential analysis* (Lyons, 1977), a linguistic theory that describes the number of common semantic components between words in different lexical relations. For example: synonymy pairs will have the biggest percentage of common semantic components, then antonymy, hyponymy, stronger contextual connection and the last will be the weaker contextual connection group. From the perspective of componential analysis, using the Spearman's correlation coefficient, we explained the significant percentage of the facilitation effect variance: ($r= -0.38$, $t(3)=2.81$, $p<0.01$). By the prediction of componential analysis, the facilitation effect will be the same no matter of the direction of the association (forward or backward). The second experiment tested this prediction.

The outcome of the second experiment indicates that the facilitation effects are not symmetrical (table 2). This could be expected on the basis of the associative norms, but it is not in accordance with previous studies. Some authors claimed that, unlike associative relations, semantic relations are symmetrical (cf. Thompson-Schill et al, 1998). However, in the present study, although the overall facilitation effect is roughly the same in both experiments (53 ms versus 54 ms) the correlation of the facilitation effects in two experiments (including semantic relations) was not significant ($r = 0.54$, $p > 0.05$).

Table 2. Average values for the stimuli in second experiment

relation type	Rt on target (ms) in neutral context (*****)	Rt on target in related context (ms)	Facilitation effect (ms) ANOVA	Facilitation effect (ms) ANCOVA
Antonymy	583	525	58	82
Hyponymy	625	583	42	32
Strong context	605	565	40	38
Weaker context	632	584	48	45
Synonymy	644	569	75	66

The results of the present study indicate that the strength of the facilitation effect depends not just upon the connection type between words (pure associative or associative and semantic connection) but also upon the relation type between them (synonymy, antonymy, hyponymy, stronger or weaker contextual connection). The fact that the type of word relation accounts for facilitation effect over and above the effect obtained by the factors known to affect priming (frequency of target and prime, rated associative connection between prime and target, associative frequency of the prime, length of the target) suggests that the semantic (lexical) relations are cognitively relevant. Furthermore, the theory of lexical semantics (i.e. componential analysis) provides good predictions of the facilitation variation among different types of relations. However, the facilitation effect is not symmetrical and varies in the direction of both the associative and semantic relation. This, on the other hand, is not in accordance with previous studies and the predictions of componential analysis.

References

Koriat, A. 1981. Semantic facilitation in lexical decision as a function of prime-target association. *Memory & Cognition, 9*(6): 587-598.

Kostić, Đ. 1999. *Frekvencijski rečnik savremenog srpskog jezika,* Beograd, Institut za eksperimentalnu fonetiku i patologiju govora i Laboratorija za eksperimentalnu psihologiju Filozofskog fakulteta u Beogradu.

Lyons, J. 1977. *Semantics*, 1-2, Cambridge: Cambridge University Press: 327-335.

Meyer, D.E. & Schvaneveldt, R.W. 1971. Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operation. *Journal of Experimental Psychology, 90*: 227-234.

Neely, J. H., Keefe, D. E., & Ross, K. L. 1989. Semantic Priming in the Lexical Decision Task: Roles of Prospective Prime-Generated Expectancies and Retrospective Semantic Matching. *Journal of Experimental Psychology: Learning Memory and Cognition, 15*(6): 1003-1019.

Piper, P., Dragićević, R., & Stefanović, M. 2005. *Asocijativni rečnik srpskoga jezika*, Beograd: Beogradska knjiga, Službeni list SCG, Filološki fakultet u Beogradu.

Rečnik MS, 1967-1976. Rečnik srpskohrvatskoga književnog jezika t. I-VI, Novi Sad: Matica srpska.

Schelton, J.R. & Martin, R.C. 1992. How semantic is automatic semantic priming? *Journal of Experimental Psychology: Learning, Memory & Cognition, 18*: 1191-1210.

Thompson-Schill, S. L., Kurtz, K. J., & Gabrieli, J. D. E. 1998. Effects of Semantic and Associative Relatedness on Automatic Priming. *Journal of Memory and Language, 38*(4): 440-458.

Elsi Kaiser
University of Southern California
elsi.kaiser@usc.edu

Looking past the pronoun

Introduction

It is widely assumed that a pronoun is preferentially interpreted as referring to whatever referent is most salient when the pronoun is encountered. On this view, information that *precedes* the pronoun plays a central role in guiding pronoun interpretation. For example, given the widespread view that subjects are by default more salient than objects (Grosz et al. 1995 and many others), as well as the claim that 'result'-connectives work against this default subject preference and focus attention on objects (e.g., Stevenson et al. 2000), a subject pronoun is more likely to refer to the subject of the preceding sentence in ex.(1a) than ex.(1b).

(1a) *Bob tickled Jim and **then** he...*
(1b) *Bob tickled Jim and **as a result** he...*

Here, we investigate effects of information not available to the processing system until *after* the pronoun has been encountered. It has been observed in previous work (especially in computational linguistics, e.g., Winograd 1972, Grosz et al. 1995, Kehler 2002 and others) that information available after the pronoun (e.g. verb semantics) may influence reference resolution. For example, work by Kehler (2002) and colleagues treats pronoun resolution as a side effect of establishing coherence relations between clauses, which is a process that makes use of both pre-pronominal and post-pronominal information. However, most existing psycholinguistic research on pronoun resolution has traditionally tended to focus on the effects of information available before the pronoun, and there has been relatively little systematic psycholinguistic investigation of what kinds of post-pronominal factors have an impact.

In this talk, we aim to contribute to our understanding of how post-pronominal information impacts reference resolution by testing whether the interpretation of sentence-initial ambiguous pronouns is influenced by the referential properties of the remainder of the sentence (see also Centering-Theoretic research by Grosz et al. 1995 and others). We also investigate how the effect of referential properties interacts with the coherence relation between two clauses (as indicated by the connectives 'and then' and 'and as a result'). This research aims to provide empirical results that can be used to enrich existing theories of reference resolution.

We take as our starting point existing psycholinguistic and cognitive psychology research which has shown that (i) referential processing imposes demands on the resources available to the human sentence processing mechanism (e.g., Warren & Gibson 2002), and that (ii) the human sentence processing mechanism (HSPM) has limited cognitive resources and thus prefers to minimize processing load whenever possible. Building on (i), it seems reasonable to hypothesize that an intransitive sentence (one argument requiring resolution) carries less processing load than a transitive

sentence (two arguments that need to be resolved). Building on (ii), we explore the Processing Cost Hypothesis which predicts that the presence/absence of subsequent referents in the rest of the clause influences whether an ambiguous subject pronoun is interpreted as referring to the preceding clause's subject or object, *with object interpretations being more likely if no further referents are mentioned in the pronoun-containing clause.* The Processing Cost Hypothesis derives this prediction from the claim that HSPM strives to minimize processing cost.

The specific prediction is generated as follows. Let us assume that, upon encountering an ambiguous pronoun, the HSPM activates both the preceding subject and object as possible antecedents, with the default subject preference modulated by the connective as shown in (2) (as predicted by work discourse connectives, e.g. Stevenson et al. 2000 and others):

(2a) *then*: subject >> object (2b) *result*: object > subject

If the Processing Cost Hypothesis is on the right track, encountering another argument later in the pronoun-containing clause increases processing load, and in response to this, to lower processing load, the HSPM gives more consideration to the default ('easy') interpretation, namely *the preceding subject*. The resulting expectation is that ambiguous pronouns (in subject position) in transitive and intransitive clauses (3a,b) will show different degrees of preference for the subject and the object of the preceding clause. Specifically, due to the HSPM striving to minimize processing load, the Processing Cost Hypothesis predicts that there should be more subject interpretations overall in transitives (3b) than intransitives (3a), and more object interpretations overall with intransitives (3a) than transitives (3b). We conducted two experiments to test this prediction.

(3a) ***X*** *verbed* ***Y*** *and {then/as a result}* ***she*** *verbed.*
(3b) ***Y*** *verbed* ***Y*** *and {then/as a result}* ***she*** *verbed the noun.*

Experiment 1

In Experiment 1, participants listened to two-clause sequences (ex.(4a-c)) and answered questions about them that probed the interpretation of the subject pronoun. We manipulated the discourse connective (*and then/and as a result*) and verb transitivity (intransitive (4a), transitive with pronominal object (4b), transitive with NP object (4c)). Nonsense words were used in place of verbs and nouns in order to factor out any effects of verb semantics in order to focus on the effects of argument frames. The sentences were spoken with neutral intonation.

(4a) *Anne tulvered Kate and {as a result/then}* ***she*** *sprelled.*
(4b) *Anne tulvered Kate and {as a result/then}* ***she*** *sprelled her.*
(4c) *Anne tulvered Kate and {as a result/then}* ***she*** *sprelled the jeg.*

Results

Participants' responses to the questions show that their interpretation of the subject pronoun is influenced by connective type and by transitivity. As predicted on the basis

of previous work, the subject pronoun is more likely to be interpreted as referring to the preceding object with 'as a result' than 'then' ('Result' conditions: about 35% subject choices on average; 'Then' conditions: >80% subject choices on average). Crucially, transitivity also has an effect: Within the 'then' and the 'result' conditions, there are significantly (p's<.05) more subject-interpretations with transitives (4b,c) than intransitives (4a), as shown in (5) below. The two types of transitives (pronominal object, (4b), and NP object, (4c)) show similar choice patterns and do not differ significantly from each other.

(5) Results: Approx. % of *subject* choices:

Result/Intransitive = 23%
Result /Transitive+noun = 42%
Result /Transitive+pronoun = 43%

Then/Intransitive = 61%
Then/Transitive+noun = 94%
Then/Transitive+pronoun = 89%

However, Exp.1 leaves open the possibility that the transitivity effect stems from the intransitives being interpreted as involving non-agentive subjects (e.g., as unaccusative verbs). Perhaps the increased number of object interpretations with intransitives results from a bias to interpret a non-agentive subject as coreferential with the preceding non-agent (i.e., the object)?

Experiment 2

Experiment 2 investigated this possibility by using real verbs in the second critical clause, including intransitive verbs with agentive subjects (unergatives, e.g. *sleep*) and intransitive verbs with non-agentive subjects (unaccusatives, e.g., *arrive*). As in Experiment 1, participants heard two-sentence sequences and responded to questions about them. The results show that unaccusatives and unergatives do not differ significantly from each other, indicating that the transitivity effect cannot be attributed to a non-agentive subject interpretation.

Discussion

Our findings highlight the importance of including the impact of post-pronominal information in theories of reference resolution. The results show that pronoun interpretation is susceptible to the referential properties of the rest of the clause – specifically, object interpretations are more likely in intransitives (i.e., if no further referents are mentioned in the pronoun-containing clause) than in transitives. A possible explanation for our finding that the *presence of subsequent arguments is correlated with an increased likelihood of subject interpretations* comes from the Processing Cost Hypothesis. Further research investigating the incremental processing load induced by sentences such as those in ex.(4) will help assess the validity of this hypothesis.

Could parallelism effects be responsible for the effects that we observed? A sizeable body of existing work (e.g. Smyth 1994) has shown that pronouns in a

particular structural position prefer antecedents realized in the same structural position (parallelism effect) – in other words, subject pronouns prefer subject antecedents and object pronouns prefer object antecedents. However, since all of our critical sentences contained subject pronouns, one possible parallelism-based prediction would be that all conditions should show equal amounts of subject preference. This, however, is not what we found, which argues against a parallelism account. Furthermore, and more crucially, it was observed that both transitive conditions (pronominal object, ex.(4b), and NP object, ex.(4c)) show an increase in subject interpretations relative to the intransitive condition, even though the second clause in the NP-object condition (4c) is not referentially parallel to the first clause. This seems to provide further evidence against a parallelism account (see also Kertz et al. 2006 for recent work suggesting that structural parallelism is not sufficient to explain patterns of reference resolution).

Conclusions

The finding that post-pronominal information has a significant effect seems to suggest that sentence-initial pronouns do not receive their final interpretation at the point at which the pronoun itself is encountered. Rather, our findings indicate that the referential properties of the remainder of the clause (i.e., whether it is transitive or intransitive) have an effect on the final interpretation assigned to subject position pronouns, possibly due to processing cost considerations. More generally, these results support the idea that psycholinguistic models of pronoun resolution will benefit from incorporating effects of post-pronominal information more fully.

References

Grosz, B. J., Joshi, A. K., & Weinstein, S. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2), 203–225.

Kertz, L. Kehler, A. & Elman, J. 2006. Evaluating a coherence-based model of pronoun interpretation. In *Proceedings of the Ambiguity in Anaphora Workshop*, ESSLLI 2006, edited by R. Artstein and M. Poesio, 49-56.

Kehler, A. 2002. *Coherence, reference, and the theory of grammar*. CSLI.

Smyth, R. H. (1994) Grammatical determinants of ambiguous pronoun resolution. *Journal of Psycholinguistic Research*, 23, 197-229.

Stevenson, R. J., Knott, A., Oberlander, J., & McDonald, S. 2000. Interpreting Pronouns and Connectives. *Language and Cognitive Processes* 15:3, 225-262.

Warren, T. & Gibson, E. 2002. The influence of referential processing on sentence complexity. *Cognition*, 85, 79-112.

Winograd, T. 1972. *Understanding natural language.* Academic Press, New York.

Liliana Martínez
Norwegian University of Science and Technology
liliana.martinez@hf.ntnu.no

Some thoughts on the semantics of non-straight paths

(based on a corpus study of Bulgarian motion verbs)

Introduction and general question

Path is an important part of the cognitive and linguistic representation of motion events. Research on conceptualization has shown that there is a conceptual distinction between straight and non-straight paths. A question which deserves further investigation is whether this distinction spills over in language in any systematic way. Is there a distinction in the overt realization of verbs encoding straight and non-straight paths? In what ways are non-straight paths encoded? The bulk of linguistic research on paths of motion is mainly about straight paths. Here I will discuss the syntactico-semantic contexts in which the most important Bulgarian verbs of non-straight motion are used, to draw a conclusion about the different ways in which they semantically represent non-straight motion.

The target verbs in this study are (1) *zavija/ zavivam* - 'turn', *izvija/ izvivam* – 'veer'; (2) *krivna/ krivvam, svurna/ svurvam/ svrushtam, svija/ svivam* – 'turn', 'take a detour'; (3) *krivolicha, lukatusha* - 'wind', 'meander'; (4) *zaobikolja/ zaobikaliam* - 'go around an obstacle'; and (5) *obikolja/ obikalyam* - 'go round'. All these verbs refer to motion along a non-straight path, but in different ways: Some of them indicate change of direction (groups 1 and 2) or multiple change of direction (group 3), others (groups 4 and 5) relate the outline of the path of motion to a Reference Object. The linguistic tradition is unanimous that recurring syntactic/ semantic contexts typical of a verb/ verb class are indicative of its semantic structure (Levin 1993, Levin & Rappaport 1995, Divjak & Gries 2006, Dimitrova-Vulchanova & Dekova 2006). The purpose of the current explorative corpus study is to map out the path-specifying context in which each of the target verbs appear. Questions which I will attempt to answer are: (A) Are there clear patterns in the types of path specification? (B) Do some of the verbs form groups based on their typical patterns? (C) What features in the semantic specification of the verbs are responsible for this? Here I discuss only spatial properties of the encoded paths. Features whose influence is important but is out of the scope of this paper are the aspectual properties of Bulgarian verbs and the clause as a whole, the boundary-crossing constraint in Bulgarian, the derivational relations between some of the target verbs, and the polysemy of the target verbs seen in connection with their different morpho-syntactic realizations. In order to be able to link the results of this study future findings about the influence of these features, all target verbs are treated as separate items in the analysis.

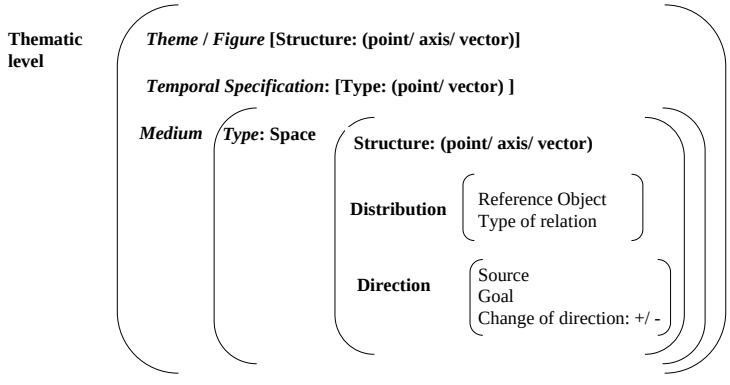
Background and hypothesis

Semantically, motion is a relation (encoded mainly in the verb) between participants (encoded in the verb or in syntactic dependents of various types). Formally, situations

are represented as functions which take a specific number of arguments belonging to specific types (Jackendoff 1983). Here I assume the decomposition approach, according to which the situation and its pertaining roles are not atomic, but can be represented as sets of many features, allowing recursion and embedding (Pustejovsky 1995, Dimitrova-Vulchanova 1996/ 99). Dependents in the clause unite with the verb by fitting into 'slots' in its semantic and syntactic matrix, which means that they must be semantically and syntactically compatible with it. Koenig et al. (2003) make the distinction between semantic arguments and syntactic dependents. Semantic arguments specify particular participants in the situation and pertain to a specific situation or type of situations. Semantic adjuncts specify the situation as a whole, and can co-occur with many situations. Thus, a set of contexts typical of a verb or set of verbs, and distinguishing it from other (sets of) verbs may be used as an indicator of common semantic properties of the verbs from the set (Levin 1993, Levin & Rappaport 1995, Divjak & Gries 2003, Dimitrova-Vulchanova & Weisgerber 2007)

The semantic properties of a situation are represented in hierarchical feature structures: with attributes taking one of a (predetermined) set of values. The value of an attribute may be *fixed* (obligatory for the verb and understood no matter whether the argument is realized overtly or not) or *specifiable* (an 'empty slot' in the situation's semantic structure which can be optionally specified further) (Dimitrova-Vulchanova 2004, 2008). A situation is specified on several levels which are independent but correspond with one another through coindexation, e.g. Jackendoff's Action Tier (involving 'social' force-dynamic) and Thematic (involving spatial relations between the participants). Only the Thematic tier is of interest for me here. A motion function (cf. Fig. 1) has two arguments: the Moving Object (called *Theme* in Gruber 1965)/ *Figure* in Talmy 1978, 1985) and *Medium* relative to which the moving or located entity is described (Dimitrova-Vulchanova 1996/ 99). Motion is a subtype of spatial relations, but differs from static situations (e.g., Location and Orientation) in its temporal specification. *Time*, *Medium* and *Figure* can be represented as projections of features along axes linked to one another by a structure-preserving bounding relation (Jackendoff 1996). In static locative situations both *Time* and *Medium* are points. In static situations involving directed paths (e.g., Orient and Extend) time is again a point, but *Medium* is projected as a vector, and thus can be further specified for its origin (Source) or end point (Goal). In situations of directed motion, both *Time* and *Medium* are projected as vectors.

Figure 1. The semantic representation of Motion and Path



Non-straight paths can also be handled under this schema: Change of direction is a subtype of directed space, and is subject to the same constraints for overt expression, e.g. the Unique Vector Constraint defined by Bohnemeyer's (2003). There are situations (both static and dynamic) in which the non-straightness of the path is not encoded in terms of change of direction, but, rather, the points of the path participate in a particular spatial configuration defined on the basis of their location with respect to a Reference Object, as in the relations specified by the prepositions *round*, and *along* (Nikanne 1990, Kray et al. 2001, Zwarts 2006). In such cases we have spatial extension without direction, for which I suggest the value *axis* can be adopted. Such paths can be specified for Reference Object and the Relation Type, however, they lack direction, and therefore cannot be specified for Direction properties. In directed paths there is both extension and direction, so they should be able to be specified for both Distribution and Direction features.

Table 1. The most common path expressions in Bulgarian

Distributed	**Directed Source**
Iz – 'throughout' ***Prez*** – 'through'/ 'across' ***Kraj/ pokraj*** – 'along (the contour of)' ***Po*** – 'along (the surface of)' ***Okolo*** – 'round' **…** **Multiple direction:** ***Nasam-natam*** – 'hither and thither' ***Naljavo-nadjasno*** – 'to the left and to the right' ***Ot*** X ***na*** X – 'from X to X' **…**	***Ot*** – 'from' ***Otkum*** – 'from the direction of' **…** **Directed Goal** ***Kum*** – 'towards' ***Na-*** + relative or ablosute direction: ***naljavo/ nastrani/ dagore/ na sever*** 'to the left'/ 'to the side'/'upwards'/ 'to the north' ***V-*** + relative direction: **v*ljavo/ vstrani*** 'to the left'/ 'to the side'/ ***V*** + NP – 'into' **…**

The occurrence of the target verbs with contexts of one or both of these types depends on the semantic specification of the verb, and the frequency of occurrence should be indicative of whether/ how each of the two properties is present in the lexical specification. If a verb combines only with path expressions of the distributed type, the path structure in its lexical specification is of type *axis*. If, however a verb combines with both directed and distributed path phrases, it is of type *vector*. If a verb (a) co-occurs with a particular type of context only optionally, and (b) this context does not refer to a participant which is 'understood' even without it being present, but(c) the context is specific for the verb, and no (group) then this context realizes overtly a feature that is lexically encoded as *specifiable*. If a verb occurs with a context fulfilling conditions (a) and (c) but is understood when omitted, this context realizes overtly a feature that is lexically encoded as *fixed*. To check this hypothesis, a corpus study of the target verbs vas conducted, following the method outlined in Divjak & Gries (2006).

Method

2626 occurrences of the target verbs taken from the Bulgarian Written Corpus (Bulgarian Academy of Sciences) and the Internet. Only dynamic motion-along-path senses of the verbs were considered. Occurrences were included on a partially random basis: A search was separately conducted or each simple verb form, and the first 20 (if found non-repeating) examples of each form were taken. Not all possible verb forms

were found, and adjectival participles and deverbal nouns were not included. Each case was coded for (1) Morpho-syntactic (aspect, transitivity, diathesis, tense, person, number) and semantic (metaphorical, fictive motion of real motion use) properties; and (2) Path-specifying syntactic dependents in the VP: (a) Direct Object (from now on DO)encoding a Reference object; and (b) path-specifying adverbial phrases. A further distinction between the following types of adverbial were made: Directed (with subtypes Source and Goal) and Distributed. The adverbial phrases extracted from the corpus include the prepositions only in their motion sense, as in their locative sense they localize the whole situation rather than its Path participant alone. In the case of preposition polysemy, different tags were used for the different meanings of a preposition. A two-dimensional crosstabulation was made for the interdependence of the target verbs with each of the contexts, and it was used as basis for a hierarchical cluster analysis.

Results

The results show that the target verbs form groups, depending on the path-specifying expressions they appear with. The verbs *Obikolja, zaobikaolja, zaobikaljam* and *obikaljam* encode distributed paths as a fixed parameter. The verbs *krivolicha* and *lukatusha* also encode distributed paths. All other target verbs express directed motion as a specifiable parameter, with a strong indication that the verbs k*rivna* and *krivvam* have at least the Source parameter *fixed.* The verbs in the directed group are very close in their meaning, but there are subtle differences which have to be determined with other means. Additional studies can be devised to check whether their differences are not in speed, degree of deviation from the original direction (cf. Klippel et al. 2005), smooth vs. sharp change (cf. van der Zee 2000), etc.

Table 2. Frequency of of occurrences of each verb with a given context. (in percent from the number of all occurrences of the verb in the data)

count	verb	distributed	directed_source	directed_goal	directed	advbl	DO
42	izvija	9,50		52,4	52,4	59,50	0,00
100	zavivam	19,00	4,00	55,00	57,00	72,00	0,00
35	izvivam	20,00		57,1	57,1	71,40	0,00
55	svurvam	18,20	10,90	60,00	69,10	85,50	0,00
142	svija	24,60	5,60	69,00	73,20	89,40	0,00
199	zavija	12,60	3,00	71,9	72,9	81,90	0,00
14	svivam	28,60		78,60	78,60	92,90	0,00
19	svrushtam	15,80	10,50	78,90	84,20	89,50	0,00
286	svurna	14,70	15,40	71,30	82,20	92,30	0,00
270	krivna	12,60	31,90	41,90	72,20	82,60	0,00
185	krivvam	8,60	40,50	38,40	75,10	81,10	0,00
251	krivolicha	52,60	2,40	14,00	15,10	59,00	0,00
204	lukatusha	59,80	2,90	12,30	12,70	64,70	0,50
256	obikaljam	35,50	3,90	6,6	6,6	35,90	53,00
222	obikolja	5,40		0,50	0,50	5,90	93,00
145	zaobikaljam	4,80		0,7	0,7	5,50	87,00
201	zaobikolja	3,50	0,50	1,5	1,5	5,00	85,00

Irine Melikishvili
Institute of General and Comparative Linguistics
Tbilisi State University
imelikishvili@yahoo.com

Functional (quantitative) hierarchies of the features voiced/voiceless and front/back

1. Phonostatistic structure of the text is determined by phonetic properties of the phonemes and their combinations. In many cases the markedness hierarchy of phonemes has conditioned character. Consonants with different degree of voicing and tension have different quantitative characteristics. In the class of labials and velars voiced and voiceless phonemes have different functional value. Phonostatistics allows to state functional hierarchies of differential features in their different combinations. In the paper the features of laryngeal articulation (voicing, aspiration and glottalization) and place of articulation (front/back) are discussed on the bases of the data of the languages of Caucasian and European linguistic areas. Two groups of regularities concerning these features are presented, one of typological and the other of universal character.

2. Phonostatistic study of the obstruents with different degrees of voicing and tension revealed their different functional value in the languages of Caucasus and Europe. Quantitative analysis of the consonant systems of 16 languages of the Caucasus (Old and Modern Georgian, Svan, Megrelian, Laz, Abkhaz, Andi, Akhvakh, Botlikh, Dido, Lak, Kubach, Tabasaran, Bats, Chechenian, Ossetic) showed their common phonostatistic characteristics: functional hierarchies of phonemic series in the classes of a) stops and b) affricates and fricatives are different in these languages. In the class of stops there is the hierarchy*: half-voiced /b d g/ > voiceless aspirated /p‘ t‘ k‘/ > glottalized /p’ t’ k’/*; in the class of affricates: *aspirated /c‘ č‘/ > glottalized /c’ č’/ > voiced /dz dž/;* in the class of fricatives: *voiceless /s š x/ > voiced /z ž γ/.* The mean quantitative characteristics of the obstruents of these 16 languages are as follows:

Stops

b	2.73	>	p‘	0.57	>	p’ 0.18
d	3.95	>	t‘	1.85	>	t’ 0.81
g	1. 69	>	k‘	1.54	>	k’ 1.23

Affricates

c‘	0.72	>	c’	0.56	>	dz 0.15
č‘	0.91	>	č’	0.42	>	dž 0.45

Fricatives

s	2.77	>	z	0.77		
š	1.49	>	ž	0.38		
x	1.93	>	γ	0.83		

This conditioned quantitative hierarchy is determined by the different degree of voicing in different classes of “Caucasian” obstruents: stops */b d g/*

are weakly voiced and the voiced affricates */dz dž/* and fricatives */z ž γ/* have high degree of voicing. It can be suggested, that the systems with analogous phonetic properties will show similar hierarchies.

Functional hierarchy in languages with the same degree of voicing in the classes of obstruents is different from the above discussed. In most of these languages the markedness hierarchy *voiceless>voiced* is attested in all classes of consonants. Mean quantitative characteristics of the stops of 13 languages (English, Dutch, French, Spanish, Portuguese, Latin, Greek, Lithuanian, Russian, Ukrainian, Bulgarian, Czech, and Hungarian) illustrate this regularity:

p 2.75 > b 1.22
t 6.36 > d 3.14
k 3.66 > g 1.07

The same relationships are in the classes of fricatives and affricates in the languages of this type.

3. The next regularity discussed in the paper has universal character. The associations of the features of the place of articulation with the features voiced/voiceless produce reversed functional hierarchies. In all types of systems *voiced labial is more frequent than the voiced velar*, but *within the class of voiceless phonemes the velar is more frequent then the labial*. So the reversed hierarchies are stated: *b>g and k>p.* This is the general relationship between all kinds of voiced and voiceless phonemes (voiced phonemes with all degrees of voicing; tense and lax phonemes; plain, aspirated and glottalized phonemes). This generalization can be illustrated by mean quantitative characteristics of the phonemes of both “Caucasian” and “European” types:

Caucasian type
b 2.73 > g 1. 69
k‘ 1.54 > p‘ 0.57
k’ 1.23 > p’ 0.18

European type
b 1.22 > g 1.07
k 3.66 > p 2.75

Consequently, the association of the features *voiced and labial can be regarded as less marked* than the association of voiced with velar and, on the other side the association of *voiceless with velar is less marked* than the association of voiceless with the labial. This generalization is in accordance with the distribution of the gaps in consonant systems (the gap of the system can be regarded as the zero point of quantitative decrease).

Elina Sellgren
University of Tampere, Finland
elina.sellgren@uta.fi

Exploring competing patterns of verb complementation: *Prevent* in the British National Corpus

The case of *prevent me from going* vs. *prevent me going*

The tight competition between the two nearly identical sentential complements of *prevent* in British English was first noted by Mair (2002) and shown to have emerged over the 20th century. Using the LOB, FLOB, Brown, and Frown corpora, Mair showed that the variant *prevent me going* (or NP-*ing*) was rare in BrE still in the 1960s, but in the 1990s it was being used at a 50:50 ratio with *prevent me from going* (or *from-ing*). NP-*ing* may remain in equal use, or eventually replace *from-ing*. In American English, only *from-ing* seems to be used. While this variation has been much researched (e.g. Van Ek (1966), Dixon (1995), Rohdenburg (1995), Mair (ibid.), Heyvaert et al. (2005), and Babováková (2005)), no clear explanations have been found for the variation. In Sellgren (2007), the phenomenon was approached by running searches in the British National Corpus (BNC) by using a search facility called the Sketch Engine.

In a pilot study, the verb forms of *prevent* seemed to favour the different variants to different degrees. This phenomenon was attributed to the Complexity principle, formulated by Rohdenburg (e.g. 1996). The principle states that in competition, the structurally more explicit variant (here *from-ing*) will be favoured in cognitively complex environments, such as passivized sentences. The less explicit variant (here NP-*ing*) will accordingly be used more often in less complex environments.

The principle is seen at work in the case of passivized examples of *prevent*, when the more explicit *from-ing* is nearly always used.[1] Different verb forms, however, hardly represent cognitively complex environments. Nevertheless, in the light of the results, there seems to be a connection between the simpler verb forms *prevent* and *prevents* and their favouring the less explicit sentential variant, NP-*ing*.

Methods

The searches were done in the whole BNC, as well as divided into the written, spoken, written-to-be-spoken parts and according to time periods (1960-1974, 1975-1984, and 1984-1995). One drawback to using the BNC as a diachronic corpus is that the earliest period only contains works of fiction, whereas the latter periods include both imaginative and informative texts.

The results were filtered in the Sketch Engine, as a lemma search for *prevent* in the BNC gives as many as 10,439 hits. Each verb form was searched separately. The concordances were filtered with a Corpus Query Language (CQL) string in order to prune out all examples of nominal complementation (of the type *I prevented the accident*), as well as either of the sentential variants. Ideally, the filtered concordances would include all and only examples of either the NP-*ing* or *from-ing* variant.

[1] In the BNC, only one example of a passivized *prevent me going* was found with a tag sequence search.

The string for *from-ing* was [word="from"][tag="V.*G"].[2] In the case of NP-*ing*, the string was [tag="N.*|PN.|DT0"][tag="V.*G"].[3] The search span was set as 1 to 7 after the key term, so that the items defined in the CQL string are found within 1 to 7 words after the key term. This span makes the manual checking of the results convenient. Searches with bigger spans did not produce a significantly higher number of additional examples.

In passive use, *from-ing* is the only choice (*I was prevented from going there* vs. **I was prevented going there*). When querying for examples of *from-ing* with *prevented*, the span of 2 to 8 was hence used to exclude passivized examples. *Prevented* in the passive is followed by *from* in position 1, except when there is a *by*-agent present. With NP-*ing*, passive examples should not occur, hence the span was set from 1 to 7 with all verb forms.

Results

In the tables below, the bottom row shows the total of examples of each variant, as well as their total percentages in relation to each other. The right-most column shows the total of examples by each verb form. Tables 1 and 2 present the distribution of the variants in the whole BNC and the written section respectively. Table 3 shows the distribution in the spoken section. Table 4 shows the distribution in the written-to-be-spoken corpus. Table 5 combines results from the subcorpora of different time periods.

In the whole BNC, the overall distribution of the variants in relation to each other is 58% for *from-ing*, and 42% for NP-*ing*. This is rather different from the 50:50 ratio found in Mair (2002). However, with the base form of *prevent* the distribution is even between the variants. With the other verb forms, the variation is strongly tilted in favour of *from-ing,* ranging from 62% to 79%.

Table 1. The whole BNC.

Verb form	*from-ing*	*%*	NP-*ing*	%	**Total**
Prevent	1399	50	1387	50	2786
Prevents	234	62	142	38	376
Preventing	325	73	122	27	447
Prevented	574	79	155	21	729
Total	2522	58	1806	42	4328

In the written section, the variation is roughly the same. This is due to the fact that the spoken section yielded remarkably few examples of *prevent* with sentential complements overall.

Table 2. The written section of the BNC.

Verb form	*from-ing*	*%*	NP-*ing*	%	**Total**
Prevent	1365	51	1337	49	2702
Prevents	222	61	140	39	362
Preventing	320	74	111	26	431
Prevented	571	79	151	21	722
Total	2478	59	1739	41	4217

[2] This string should include only those examples of *prevent* where it is followed by the word *from* as well as an *-ing* form.

[3] Here common nouns are represented by "N", demonstrative pronouns by "DT0", and all other pronouns by "PN".

In the spoken section, only 121 examples of *prevent* altogether were found with sentential complements. The distribution is strikingly different from the written section, and the BNC as a whole: NP-*ing* is favored in 60% of the cases. With the base form of *prevent*, the ratio is the same. The other verb forms were present in such few numbers that no conclusions can be drawn.

Table 3. The spoken section of the BNC.

Verb form	*from-ing*	*%*	NP-*ing*	%	**Total**
Prevent	34	40	50	60	84
Prevents	12	86	2	14	14
Preventing	5	31	11	69	16
Prevented	3	43	4	57	7
Total	54	40	67	60	121

In the written-to-be-spoken section, NP-*ing* is favored overall even more prominently than in the spoken section: it is used in 68% of the examples. Even though the total number of examples is rather small (53), the tendency of the written-to-be-spoken texts to strongly favor NP-*ing* seems clear. The distribution seems to mimic that found in the spoken section of the BNC, in great contrast with the distribution found in the written section and the whole BNC.

Table 4. The written-to-be-spoken section of the BNC.

Verb form	*from*	*%*	NP-*ing*	%	**Total**
Prevent	7	17	34	83	41
Prevents	1	50	1	50	2
Preventing	4	57	3	43	7
Prevented	2	67	1	33	3
Total	14	32	39	68	53

Comparing the results obtained by Mair (2002) and those obtained in this study show how different corpora can give different pictures on competition between complementation variants. Mair's diachronic perspective, albeit limited to the 20th century, demonstrated that NP-*ing* has become a serious competitor to *from-ing* only very recently. This view is corroborated by the results from the diachronically divided subcorpora in the BNC.

Table 5. Subcorpora of different time periods in the BNC.

1960-1974	*from*	*%*	NP-*ing*	%	**Total**
Prevent	27	64	15	36	42
Prevents	7	100	-	0	7
Preventing	11	100	-	0	11
Prevented	22	96	1	4	23
Total	67	81	16	19	83
1975-1984	*from*	*%*	NP-*ing*	%	**Total**
Prevent	68	55	55	45	123
Prevents	23	85	4	15	27
Preventing	19	86	3	14	22
Prevented	38	92	3	8	41
Total	148	69	65	31	213

1985-1995	*from*	%	NP-*ing*	%	**Total**
Prevent	1304	50	1317	50	2621
Prevents	204	60	138	40	342
Preventing	295	71	119	29	414
Prevented	514	77	151	23	665
Total	2317	57	1725	43	4042

Table 5 shows how NP-*ing* was marginal in the 1960-1975 period: it represents only 19% of all examples. Significantly, all examples but one occurred with the base form. This variant is slightly more common in the 1975-1984 period with 31%. Again, other verb forms than the base form are scarcely found at all with NP-*ing*. In the last period, 1985-1995, NP-*ing* is found increasingly also with *prevents*, in addition to *prevent*. Apparently NP-*ing* has evolved into an equal competitor to *from-ing* mostly through the base form of *prevent*.

Conclusion

The exploratory searches in the BNC with the Sketch Engine have given interesting pointers to future studies on *prevent*. The competition between NP-*ing* and *from-ing* may indeed be a case of the advance of one at the expense of the other. NP-*ing* is not necessarily in perfectly equal variation with *from-ing* today, as Mair suggested. The diachronic subcorpora showed that NP-*ing* has advanced first and foremost with the base form, and with this verb form it is in equal variation with *from-ing* both in the whole BNC and the written section. The Complexity principle may cause the less explicit variant NP-*ing* to be used more often used with the simplest verb form, *prevent*, and also with *prevents* in the latest time period.

In the spoken section, the tables are turned: NP-*ing* has a lead over *from-ing* with 60%. It could be hypothesized in the spirit of Mair that spoken texts are the most "advanced" as regards the idea that NP-*ing* will eventually overcome *from-ing*. However, the religious use of *from-ing* with passivized examples makes one wonder whether NP-*ing* can truly become the sole option as a sentential complement of *prevent*.

References

Babováková, Petra. (2005) The Complements of *prevent*. A Master's Thesis. Tampere: Tampere University Press.

Dixon, R.M.W. (1995)"Complement clauses and complementation strategies" in F.R. Palmer (ed.), *Grammar and Meaning: Essays in honour of Sir John Lyons*. Cambridge: University Press.

Heyvaert, L., Rogiers, H. and Vermeylen, N. (2005) "Pronominal Determiners in Gerundive Nominalization: A "Case" Study". *English Studies* Vol.86, February 2005: 71-88.

Mair, C. (2002) "Three Changing Patterns of Verb Complementation in Late Modern English". *English Language and Linguistics* 6: 105-132.

Rohdenburg, G. (1995) "Betrachtungen zum auf- und abstieg einiger präpositionaler konstruktionen im englischen". *Nowele* 26: 67-123.

---------- (1996) "Cognitive Complexity and increased grammatical explicitness in English". *Cognitive Linguistics* 7: 149-82.

Sellgren, E. (2007) A Corpus-based study of the complements of *prevent* in the 18th, 19th and 20th centuries. Tampere: Tampere University Press.